Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Mining and Machine Learning

Arthur Zimek

University of Southern Denmark

DM566/DM868/DM870, spring 2019

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outline

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outline

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Tabulæ Rudolphinæ

(observations by Tycho Brahe and Johannes Kepler)

Pars Secunda. 63

Tabula Æquationum MARTIS.

Anomalia Eccentri. Cum æquatio nis parte phys	Interco-lumnium, Cum Log-arithmo.	Anomalia coæquata.	Intervallũ Cum Loga-rithmo		Anomalia Eccentri, Cum æquatio nis parte phys	Interco-lumnium, Cum Log-arithmo.	Anomalia coæquata.	Intervallũ Cum Loga-rithmo	
60 4.35.50	9690 0.54.27	55.30. 0	159407 46630	21	90 5.18.30	560 0.59.40	84.41. 3	152350 42101	27
61 4.38.34	9430 0.54.36	56.27. 7	159193 46495	21	91 5.18.27	230 0.59.52	85.40.52	152104 41939	27
62 4.41.13	9160 0.54.45	57.24.20	158977 46359	22	92 5.18.18	100 1. 0. 4	86.40.46	151857 41777	27
63 4.43.47	8890 0.54.54	58.21.37	158758 46221	22	93 5.18. 3	430 1. 0.16	87.40.46	151611 41615	27
64 4.46.15	8610 0.55. 3	59.19. 0	158537 46082	22	94 5.17.43	760 1. 0.27	88.40.52	151366 41453	27
65 4.48.39	8320 0.55.13	60.16.27	158315 45941	23	95 5.17.17	1080 1. 0.39	89.41. 3	151121 41291	27
66 4.50.58	8030 0.55.22	61.13.59	158090 45799	23	96 5.16.45	1410 1. 0.51	90.41.20	150875 41129	27
67 4.53.11	7730 0.55.32	62.11.36	157864 45656	23	97 5.16. 7	1730 1. 1. 3	91.41.43	150630 40966	27
68 4.55.18	7420 0.55.42	63. 9.18	157637 45512	23	98 5.15.24	2060 1. 1.15	92.42.11	150386 40804	27
69 4.57.20	7120 0.55.53	64. 7. 5	157408 45367	24	99 5.14.35	2380 1. 1.27	93.42.44	150142 40642	27
70 4.59.17	6810 0.56. 3	65. 4.57	157177 45221	24	100 5.13.39	2710 1. 1.39	94.43.23	149900 40480	27
71 5. 1. 9	6510 0.56.13	66. 2.54	156944 45073	24	101 5.12.40	3030 1. 1.51	95.44. 7	149658 40319	27
72 5. 2.55	6210 0.56.23	67. 0.57	156711 44924	24	102 5.11.33	3350 1. 2. 3	96.44.57	149416 40157	27
73 5. 4.35	5900 0.56.34	67.59. 4	156477 44774	25	103 5.10.21	3670 1. 2.14	97.45.55	149175 39996	27
74 5. 6. 9	5600 0.56.44	68.57.16	156241 44623	25	104 5. 9. 3	3980 1. 2.26	98.46.59	148935 39835	27
75 5. 7.39	5300 0.56.54	69.55.34	156003 44471	25	105 5. 7.39	4300 1. 2.38	99.48. 9	148697 39674	27
76 5. 9. 3	5000 0.57. 4	70.53.57	155765 44318	25	106 5. 6. 9	4620 1. 2.50	100.49.25	148461 39514	27
77 5.10.21	4690 0.57.15	71.52.26	155526 44164	25	107 5. 4.35	4930 1. 3. 2	101.50.46	148225 39355	27
78 5.11.33	4390 0.57.25	72.51. 0	155285 44009	26	108 5. 2.55	5240 1. 3.14	102.52.14	147990 39196	27
79 5.12.39	4090 0.57.36	73.49.40	155043 43853	26	109 5. 1. 9	5550 1. 3.25	103.53.47	147756 39038	27
80 5.13.40	3790 0.57.46	74.48.25	154800 43696	26	110 4.59.17	5860 1. 3.37	104.55.27	147523 38881	27
81 5.14.35	3480 0.57.57	75.47.15	154557 43539	26	111 4.57.20	6180 1. 3.50	105.57.13	147292 38725	27
82 5.15.24	3170 0.58. 8	76.46.11	154314 43381	26	112 4.55.18	6490 1. 4. 2	106.59. 4	147063 38569	27
83	2860		154070		113	6800		146826	

Kepler's laws of planetary motion

Data Science

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Data Science

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

- “data science” is a vaguely defined concept

Data Science

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

- “data science” is a vaguely defined concept
  - meaning A: doing science based on (big) data

Data Science

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

- “data science” is a vaguely defined concept
  - meaning A: doing science based on (big) data
  - meaning B: the science of working with data

Data Science in Venn Diagrams

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Science in Venn Diagrams

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Science in Venn Diagrams

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Science in Venn Diagrams

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

Entertaining post on attempts to describe data science by Venn diagrams:

- `http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html`

(The diagrams shown on the previous slide were taken from there.)

Data Mining and Related Areas

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Mining tries to combine the strengths of many domains.

Data Mining and Related Areas

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data Mining tries to combine the strengths of many domains.

Understand Methods for Analyzing Data

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Illustration by David Parkins (detail).

Source: Silberzahn and Uhlmann [2015].

- data science: learning from data, finding patterns in data, understanding databases

Understand Methods for Analyzing Data

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Illustration by David Parkins (detail).

Source: Silberzahn and Uhlmann [2015].

- data science: learning from data, finding patterns in data, understanding databases
- data mining/machine learning: computational *methods* for learning from data

Understand Methods for Analyzing Data

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Illustration by David Parkins.

Source: Silberzahn and Uhlmann [2015].

- data science: learning from data, finding patterns in data, understanding databases
- data mining/machine learning: computational *methods* for learning from data
- different methods deliver different pictures of the data

Understand Methods for Analyzing Data

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Illustration by David Parkins.

Source: Silberzahn and Uhlmann [2015].

- data science: learning from data, finding patterns in data, understanding databases
- data mining/machine learning: computational *methods* for learning from data
- different methods deliver different pictures of the data
- this course should help:
  - to learn about data mining/ machine learning methods
  - to understand their characteristics
  - to apply them correctly
  - to derive meaningful results

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Motivation: Knowledge Discovery from Data

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- huge amount of data collected in various application domains
- manual analysis?

Definition: Knowledge Discovery from Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

"KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." [Fayyad et al., 1996]

- *data*: set of facts (e.g., entries in a database)
- *pattern*: expression in some language to describe a data subset (e.g., mathematical model)
- *process*: can involve several steps or iterations
- *nontrivial*: more complex than search, inference, simple aggregations
- *valid*: applicable to new data with a certain degree of reliability
- *novel*: for the system, better: for the user
- *potentially useful*: beneficial for user of application
- *ultimately understandable*: if not immediately then given some post processing

understandability $\Leftrightarrow$ simplicity?

(validity, novelty, usefulness, simplicity) $\Leftrightarrow$ "interestingness"

The KDD process model

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

KDD process model (cf. Fayyad et al. [1996])

- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - discretize numeric attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Example: Clustering

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Outlier

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Classification

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Regression

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Frequent Patterns

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

a,**b,c,d**,e
b,c,d
a**,b,c,d**
c,d,f
a,**b,c,d**,e
a,c,d
a,c,e,f
c,d,e,f
a,**b,c,d**,f
a,b,e,f

In 5 out of 10 (50%) cases,

b,c,d

occur together.

In 5 cases we have **b,c**, and in all those 5 cases we also have **d**:

Rule with 100% confidence:

If **b,c** *are in the set, then also* **d** *is in the set.*

Data Mining Techniques: From Data to Knowledge

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data	Method	Knowledge
	clustering	customer profile
	outlier detection	fraud
	frequent pattern mining	product groups
	classification	star type

Patterns and Models

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- *predictive models*:
  A predictive model should describe (known) data in a way suitable to make predictions on unknown data.
- *descriptive models*:
  A model should provide insight in the data, help to understand the underlying structure, correlations, properties etc.
- A data mining model can be both, descriptive and predictive.
- Based on both could be a *prescriptive model*:
  Learn from the data for the future: a model to tell you what to change, or how to implement details (typically related to optimization problems).

Data Mining Techniques: Categories of Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- *supervised* (e.g., classification, regression, outlier detection):
  An in general unknown attribute is learned, based on examples (training data) where the attribute is known. Emphasis on predictive modeling.
- *unsupervised* (e.g., clustering, outlier detection, association rules)
  The data are distinguished/organized in different groups without previously known examples. Emphasis on descriptive modeling.
- *semi-supervised* (e.g., clustering, outlier detection)
  Semi-supervised techniques are guided by using *some* information (e.g., only one class is known, or some constraints restrict clustering results)

Note that:

KDD is targeted from very different perspectives, such as researchers in statistics, in machine learning, in databases, and in applied areas such as bioinformatics or economy. Thus, terminology is not always used in the same way.

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The KDD process model

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

KDD process model (cf. Fayyad et al. [1996])

- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - discretize numeric attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Preprocessing

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data may be

- noisy in the sense of containing errors, outliers
- noisy in the sense of containing lots of irrelevant information
- incomplete (e.g., missing values, missing attributes, that would have been particularly interesting for some given task)
- inconsistent (e.g., different scaling in student evaluation sheets for different cohorts, different questions in questionaires for different universities)

Typical Preprocessing Tasks

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- data cleaning: impute missing values, smoothing of noisy values, identify or remove outliers, resolve inconsistencies
- data integration: combination from different data sources: entity identification, value resolution
- data reduction: elimination of duplicates

Note that:

Many of these operations do actually change the data, based on some assumptions – handle with care if these assumptions are not explicit!

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The KDD process model

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

KDD process model (cf. Fayyad et al. [1996])

- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - discretize numeric attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Data Descriptions

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Many techniques work on feature attributes (feature vectors).
- Other techniques work directly on complex data such as text, sets, graphs.
- If we are to perform data mining on some complex objects, it is an important preprocessing step to derive meaningful features to describe these objects.

Similarity

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Similarity (as given by some distance measure) is a central concept in data mining, e.g.:
  - clustering: group similar objects in the same cluster, separate dissimilar objects to different clusters
  - outlier detection: identify objects that are dissimilar (by some characteristic) from most other objects
- definition of a suitable distance measure is often crucial for deriving a meaningful solution in the data mining task
  - images
  - CAD objects
  - proteins
  - texts
  - ...

Deriving Features from Complex Objects

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- image database
  ↓
  color histograms
- gene database
  ↓
  expression levels
- text database
  ↓
  word counts

Data	25
Mining	15
Feature	12
Object	7
...	

Note that:

Data mining methods work on the derived feature space no matter the original nature of the object – thus the mapping to a meaningful feature space is of paramount importance.

Typical Transformation Tasks

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- scale, normalize, generalize (e.g., by concept hierarchy)
- data reduction: aggregation, feature combination, dimensionality reduction
- derive new features

Note that:

Many of these operations do actually change the data, based on some assumptions – handle with care if these assumptions are not explicit!

Representation Matters

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure adapted from Figure 1.1, Goodfellow et al. [2016].

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The KDD process model

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

KDD process model (cf. Fayyad et al. [1996])

- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - discretize numeric attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Evaluation

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

"KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." [Fayyad et al., 1996]

- The challenge of evaluation: to quantify how *valid, novel, useful and understandable* the derived patterns and models are.
  - Evaluation of descriptive models (clustering)?
  - Evaluation of predictive models (classification)?

Note that:

There is something like a 'Heisenbergian uncertainty relation' between validity and novelty.

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Resources

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- overview on free and commercial tools:
  `http://www.kdnuggets.com/software/suites.html`
- `http://www.kdnuggets.com/2015/03/machine-learning-table-elements.html`
- commercial tools offered by database companies (such as Oracle, IBM, Microsoft)
- some free / open source tools:

WEKA The University of Waikato

Environment for DeveLoping KDD-Applications Supported by Index-Structures

SciPy.org

Books (General Textbooks)

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. 3rd edition, Morgan Kaufmann, 2012
- Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, 2006
- Mitchell, T. M.: Machine Learning. McGraw-Hill, 1997
- Witten, I. H., Frank, E., Hall, M. A.: Data Mining: Practical Machine Learning Tools and Techniques. 3rd edition, Morgan Kaufmann, 2011
- Zaki, M. J ., Meira Jr., W.: Data Mining and Analysis. Cambridge, 2014

None is required, all might be helpful.

See also information on e-learn or find them on the semester shelf.

Course Material

DM566/DM868/DM870
Arthur Zimek
Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

On e-learn you find furthermore:

- these lecture slides (updated over time)
- weekly exercises (to be prepared before the tutorial!)
- data sets we use in the lecture and in exercises
- later on: suggestions for practical projects and presentation topics
  - master level students should do a presentation, possibly in groups, based on such a project or on scientific literature (we have to agree on the topic)
  - bachelor level students can do that, if they want (recommended!)

Outline

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
- Data Science
- Knowledge Discovery from Data
- Data Mining Methods
- Preprocessing
- Transformation
- Evaluation
- Material
- Summary

Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Data Science
Knowledge Discovery from Data
Data Mining Methods
Preprocessing
Transformation
Evaluation
Material
Summary
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *steps for analyzing data (KDD process model)*
- *“data science”, “data mining”, “machine learning”*
- *descriptive and predictive models*
- *supervised, unsupervised, semi-supervised*
- *paradigmatic data mining methods: clustering, outlier detection, classification, regression, frequent pattern mining and association rules*
- *typical tasks of preprocessing*
- *data transformation: data representation and similarity measures (motivating examples)*
- *evaluation: the fundamental problem of measuring validity and novelty at the same time*

Outline

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Tan et al. [2006], Chapter 6.*
- *Han et al. [2011], Chapter 6.*
- *Zaki and Meira Jr. [2014], Chapters 8+9.*
- *Advanced topics: Aggarwal and Han [2014].*

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Definition of Sets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A set is a collection of objects (e.g., $S = \{1, 4, 9\}$).

- The objects are said to be elements of the set (e.g., $1 \in S, 4 \in S, 9 \in S$). Each element is *unique*.

We can define sets

extensionally: by enumerating the elements that define the set (e.g., $S = \{1, 4, 9\}$).

intensionally: by characterizing the elements of the set,

- describing what condition (the "*characteristic function*" of the set) holds for all the elements and only for the elements of the set (e.g. $S = \{x | \sqrt{x} \in \mathbb{N}$ *AND* $x < 15\}$ — read '|' as 'for which holds' or 'such that').
- The intensional definition typically resorts to a domain over which the set is defined (here: $\mathbb{N}$).

Properties of Sets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A set can be finite (e.g., $S = \{1, 2, 3, \ldots, n\}$: $1 \in S, 2 \in S, 3 \in S, \ldots, n \in S$). In this case, the set has a finite size (*cardinality*), that is the number of elements of the set (e.g., $|S| = n$).
- A set can be countably infinite. Example:

$$\begin{aligned} S &= \{1, -1, 3, -3, 5, -5, \ldots\} \\ &= \{x | x = 2k + 1 \ \textit{OR}\ x = -2k + 1, k \in \mathbb{N}_0\} \end{aligned}$$

- A set can be uncountable (e.g., $\mathbb{R}$).
- A set can be empty: $S = \{\} = \emptyset$. $|S| = 0$.

Notations

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- logical operators:
  - $\vee$: or
  - $\wedge$: and
  - $\neg$: not
  - $\not x$: negation of x (e.g., $\neq, \notin$)
  - $A \Rightarrow B$: if A, then B (short for $\neg(A \wedge \neg B)$)
  - $A \Leftarrow B$: if B, then A (short for $\neg(B \wedge \neg A)$)
  - $A \iff B$: $A \Rightarrow B \wedge A \Leftarrow B$ ("iff", read: "if and only if", also: $\equiv$, read: "is equivalent to")
- $\exists x : p(x)$, means: there exists some x such that $p(x)$
- $\forall x : p(x)$, means: for all x holds $p(x)$
- subset: $T \subseteq S \equiv \forall x \in T : x \in S$
- proper subset:
  $T \subset S \equiv (\forall x \in T : x \in S) \wedge (\exists x \in S : x \notin T)$

Algebra of Sets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

An algebra is defined over a base set Ω, all sets involved in the algebra are subsets of Ω.

basic operations for $S, T \subseteq \Omega$:

- union $S \cup T \equiv \{x | x \in S \vee x \in T\}$
- intersection $S \cap T \equiv \{x | x \in S \wedge x \in T\}$
- complement $\bar{S} \equiv S^C \equiv \{x | x \notin S\}$
- difference $S \setminus T \equiv \{x | x \in S \wedge x \notin T\}$
- product $S \times T \equiv \{(x, y) | x \in S \wedge y \in T\}$

De Morgan's laws :

- $(S \cup T)^C = S^C \cap T^C$
- $(S \cap T)^C = S^C \cup T^C$
- $\left(\bigcup_{i \in I} S_i\right)^C = \bigcap_{i \in I} (S_i)^C$
- $\left(\bigcap_{i \in I} S_i\right)^C = \bigcup_{i \in I} (S_i)^C$

Powerset $\mathcal{P}(S) \equiv \wp(S) \equiv 2^S \equiv \{T | T \subseteq S\}$

Tupel and Relations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

n-tupel result of the (Cartesian) product of n sets:

$$S_1 \times S_2 \times S_3 \times \ldots \times S_n =$$
$$\{(a_1, a_2, a_3, \ldots a_n) | a_1 \in S_1, a_2 \in S_2, a_3 \in S_3, \ldots, a_n \in S_n\}$$

If $S_1 = S_2 = S_3 = \ldots = S_n$, we write:

$$S^n \equiv S_1 \times S_2 \times S_3 \times \ldots \times S_n$$

n-ary relation R is a set of n-tupels:

$$R(x_1, \ldots, x_n) \equiv (x_1, \ldots, x_n) \in R$$

characteristic function of $R \subseteq S_1 \times \ldots \times S_n$ (can also be seen as "predicate": "R is true/false")

$$S_1 \times \ldots \times S_n \to \{\text{true}, \text{false}\}$$

$$t \mapsto \begin{cases} \text{true} & \text{if } t \in R \\ \text{false} & \text{otherwise} \end{cases}$$

Properties of Homogeneous Relations

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Relation $R \subseteq S_1 \times S_2 \times S_3 \times \ldots \times S_n$ is homogeneous iff $S_1 = S_2 = S_3 = \ldots = S_n$.

A binary homogeneous relation $R \subseteq S \times S$ is

reflexive iff $\forall x \in S : (x, x) \in R$

symmetric iff $\forall x, y \in S : (x, y) \in R \Rightarrow (y, x) \in R$

antisymmetric iff $\forall x, y \in S : (x, y) \in R \wedge (y, x) \in R \Rightarrow x = y$

transitive iff $\forall x, y, z \in S : (x, y) \in R \wedge (y, z) \in R \Rightarrow (x, z) \in R$

total iff $\forall x, y \in S : (x, y) \in R \vee (y, x) \in R$

Examples:

- '$<$' $\subseteq \mathbb{N} \times \mathbb{N}$: $(1, 2) \in$ '$<$'
  (more customary is the infix notation $1 < 2$)
- '$=$' $\subseteq \mathbb{N} \times \mathbb{N}$: $(361, 361) \in$ '$=$' or $361 = 361$
- '$\subset$' $\subseteq \wp(S)^2$

Partial and Total Order

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A relation $R \in S \times S$ can be an order:

R is a partial order if R is antisymmetric and transitive.

R is a total order if R is a partial order and is total.

R is a strict order if R is a (partial or total) order and is not reflexive

Are these relations orders? Which kind of order?

- alphanumeric sorting of strings?
- ‘ $<$ ’ $\subseteq \mathbb{N} \times \mathbb{N}$
- ‘ $\leq$ ’ $\subseteq \mathbb{N} \times \mathbb{N}$
- ‘ $\subset$ ’ $\subseteq \wp(S)^2$
- ‘ $\subseteq$ ’ $\subseteq \wp(S)^2$

Lattice of Subsets

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Functions

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A *function* is a mapping from some set (the domain) to some set (the image).
- We can see functions as relations with particular properties: a *univalent* (or *right-unique*) relation over domain and image.
- Formally, a function f is a binary relation over $D \times I$: $f \subseteq D \times I$, for which holds:

$$(x, y) \in f \wedge (x, z) \in f \Rightarrow y = z$$

  i.e., for each $d \in D$, f maps to at most one $i \in I$.
- Notation:

$$(x, y) \in f \iff y = f(x) \iff f(x) = y \iff f : x \mapsto y$$

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Motivation

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

on the webpage of some online bookseller:

Frequently Bought Together

Total price: $188.39

Add all three to Cart

Add all three to List

- This item: Data Mining and Analysis: Fundamental Concepts and Algorithms by Mohammed J. Zaki Hardcover $67.89
- Data Mining: The Textbook by Charu C. Aggarwal Hardcover $85.49
- Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Foster Provost Paperback $35.01

Customers Who Bought This Item Also Bought

Page 1 of 17

Data Mining: The Textbook › Charu C. Aggarwal ★★★★★ 8

The Elements of Statistical Learning: Data Mining, Inference, and…

Data Science for Business: What You Need to Know about Data Mining and…

Applied Predictive Modeling › Max Kuhn

Learning From Data › Yaser S. Abu-Mostafa ★★★★½ 108

Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann…

Market Basket Analysis

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

application examples:

- layout of supermarket: optimize the arrangement of items bought together
- online seller: recommend related items
- cross-marketing, add-on sales, targeted attached mailings

read about the infamous example of beer and diapers:

`http://www.theregister.co.uk/2006/08/15/beer_diapers/`

Frequent Patterns

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Transaction database:

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
⋮
⋮
```

If we observe patterns, can we conclude on associations between items?

Associations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

a,**b,c,d**,e
b,c,d
a,**b,c,d**
c,d,f
a,**b,c,d**,e
a,c,d
a,c,e,f
c,d,e,f
a,**b,c,d**,f
a,b,e,f

In 5 out of 10 (50%) cases,

b,c,d

occur together.

In 5 cases we have **b,c**, and in all those 5 cases we also have **d**:

Rule with 100% confidence:

If **b,c** *are in the set, then also* **d** *is in the set.*

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Association Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

items bread, butter, eggs, milk etc.

transaction a set of items

rule $L \Rightarrow R$, $L, R \subseteq$ items, $L \cap R = \emptyset$

- **L** *left-hand-side* or *antecedent*
- **R** *right-hand-side* or *consequent*

- {butter, flour} $\Rightarrow$ {milk}
- {sugar} $\Rightarrow$ {butter}

Definition: Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

items: $I = \{i_1, \ldots, i_m\}$ a set of literals (e.g., items in a shop)

itemset: $X \subseteq I$ (e.g., the items in a basket)

transaction: $T = (tid, X_{tid})$ designates a specific itemset

transaction database $\mathcal{D}$: a set of transactions

order: items in an itemset are ordered by some strict total order (e.g., alphabetical order of the literals), i.e.: $X = (x_1, x_2, \ldots, x_k) \Rightarrow x_1 < x_2 < \ldots < x_k$

length of an itemset: number of elements contained in the itemset

k-itemset: an itemset of length k (e.g., T1 is a 4-itemset, T4 is a 1-itemset)

Definition: Cover, Support, Frequency

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

cover of an itemset: set of all transactions that contain the itemset: $\mathrm{cover}(X) = \{(tid, X_{tid}) | (tid, X_{tid}) \in \mathcal{D} \wedge X \subseteq X_{tid}\}$

support of an itemset: the support s of an itemset X ($s(X)$) is the number of transactions containing X (i.e., the size of the cover set): $s(X) = |\mathrm{cover}(X)|$

frequency of an itemset: the frequency of an itemset X is its support relative to the database size $f(X) = \frac{s(X)}{|\mathcal{D}|}$

frequent itemset: given some support threshold σ, an itemset X is frequent (w.r.t. σ) iff: $s(X) \geq \sigma$ or equivalently $f(X) \geq \frac{\sigma}{|\mathcal{D}|}$

Problem 1: Frequent Itemset Mining

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ

Find all frequent itemsets in $\mathcal{D}$, i.e., $\{X | X \subseteq I \wedge s(X) \geq \sigma\}$

example: which itemsets are frequent with $\sigma = 3$ in $\mathcal{D}$:

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Definition: Association Rule

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in $\mathcal{D}$ is the support of the union of its components: $s(X \Rightarrow Y) = s(X \cup Y)$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Problem 2: Association Rule Mining

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in $\mathcal{D}$ with a support of at least σ and a confidence of at least c, i.e.:

$$\{X \Rightarrow Y | s(X \Rightarrow Y) \geq \sigma \wedge \mathrm{conf}(X \Rightarrow Y) \geq c\}$$

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Problem 1 $\subset$ Problem 2

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Problem 1 is part of Problem 2:

- Itemset A is frequent w.r.t. σ
- Given
  - $A = X \cup Y$
  - $X \cap Y = \emptyset$
  - $X = A \setminus Y$
  - $Y = A \setminus X$
- '$X \Rightarrow Y$' is frequent w.r.t. σ

Two-step approach:

1. find all frequent itemsets w.r.t. σ
2. generate rules with confidence $\geq c$ from each frequent itemset, where each rule is a binary partition of the itemset

Find Frequent Itemsets: Naïve Algorithm (Brute Force)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Each possible itemset is a *candidate* for being a *frequent* itemset.
- We need to check the database to count if the itemset is actually frequent.
- Complexity roughly amounts to:
  number of candidates $\times$ number of transactions
- How many candidates do we have, given n items?

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Lattice of Itemsets

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

TID	A	B	C	D	E
1	0	1	0	0	0
2	1	0	1	1	1
3	1	1	1	0	1
4	0	0	1	1	0
5	1	1	1	1	1
6	1	0	1	1	1
7	1	1	0	0	0

Monotonicity and Anti-Monotonicity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 4$:

Monotonicity and Anti-Monotonicity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 4$:

Monotonicity and Anti-Monotonicity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 3$:

Monotonicity and Anti-Monotonicity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 3$:

Positive and Negative Border

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Due to the anti-monotonicity, we can summarize solutions by their border in the lattice:

- An itemset X belongs to the border (w.r.t. some σ), if:
  - $\forall Y \subset X : Y$ is frequent (w.r.t. σ)
  - $\forall Z \supset X : Z$ is not frequent (w.r.t. σ)
- positive border: X itself is frequent (also: "X is a maximal frequent itemset")
- negative border: X itself is not frequent

Maximal frequent itemsets can be used as a condensed representation of a solution, as all frequent itemsets can be derived from the maximal frequent itemsets.

Note that:

The anti-monotonicity property of support is also called downward-closure property.

Maximal Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$

Maximal Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$

Closed (Frequent) Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

closed itemset: An itemset X is *closed* if none of its immediate supersets has exactly the same support as X.

closed frequent itemset: An itemset is a closed frequent itemset (w.r.t. some σ) if it is a closed itemset and is frequent (w.r.t. σ).

Closed frequent itemsets represent a solution (all frequent itemsets) *and* their support.

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Apriori Algorithm [Srikant and Agrawal, 1996]: Idea

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

1. find frequent 1-itemsets first, then 2-itemsets, 3-itemsets etc. (breadth-first search in the lattice)
2. for finding $(k+1)$-itemsets C_{k+1}: consider only those as candidates, where *all* k-itemsets $C_k \subset C_{k+1}$ are frequent
3. count frequency of all k-itemset candidates in a single database scan (hashing of the candidate itemsets)

Apriori Algorithm [Srikant and Agrawal, 1996]: Pseudo Code

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

C_k: k-itemset candidates, S_k: frequent k-itemsets (solution)

Algorithm 2.1 (Apriori [Srikant and Agrawal, 1996])

```
Apriori(I, D, σ)
  S_1 = {frequent 1-itemsets};
  k = 2;
  while S_{k-1} ≠ ∅ do
    C_k = AprioriGenerateCandidates(S_{k-1});
    for each transaction T ∈ D do
      C_T = {c ∈ C_k | c ⊆ T};
      for each c ∈ C_T do
        c.count++;
    S_k = {c ∈ C_k | c.count ≥ σ};
    k++;
  return ∪_k S_k;
```

Apriori Candidate Generation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 2.2 (AprioriGenerateCandidates(S_{k-1}))

1. *join:*
   two frequent $(k-1)$*-itemsets* $p, q \in S_{k-1}$ *are joined if they are identical in the* ***first*** *(order!)* $k-2$ *items:*

$$\begin{array}{lll} p \in S_{k-1}: & (\underline{A,B,}C) & \\ q \in S_{k-1}: & (\underline{A,B,}D) & \\ \hline \Rightarrow & (A,B,C,D) & \in C_k \end{array}$$

2. *pruning:*
   remove all k*-itemsets from* C_k *that contain any* $(k-1)$*-itemset* $\notin S_{k-1}$

example: $S_3 = \{(1,2,3),(1,2,4),(1,3,4),(1,3,5),(2,3,4)\}$

1. join: $C_4 = \{(1,2,3,4),(1,3,4,5)\}$
2. pruning: remove $(1,3,4,5)$

result: $C_4 = \{(1,2,3,4)\}$

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Apriori Example: Find Frequent Itemsets X with $f(X) \geq 0.3$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	
B	
C	
D	
E	
F	
G	
H	
I	
J	
K	
L	

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	
B	1
C	
D	
E	1
F	
G	1
H	1
I	
J	
K	
L	

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	1
B	2
C	1
D	
E	2
F	
G	2
H	2
I	
J	
K	
L	

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	2
B	3
C	2
D	
E	3
F	1
G	2
H	3
I	
J	
K	
L	

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	2
B	4
C	3
D	1
E	4
F	2
G	3
H	4
I	
J	
K	
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	3
B	5
C	3
D	1
E	5
F	2
G	3
H	5
I	
J	
K	1
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	3
B	6
C	3
D	1
E	6
F	3
G	4
H	6
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	4
B	7
C	3
D	2
E	6
F	3
G	5
H	7
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	8
C	3
D	3
E	6
F	3
G	6
H	7
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	9
C	3
D	4
E	6
F	4
G	7
H	7
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	9
C	4
D	4
E	7
F	5
G	7
H	7
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	6
B	9
C	5
D	4
E	8
F	6
G	7
H	8
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8
I	1
J	
K	2
L	1

C1 and S1

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8
I	~~1~~
J	~~0~~
K	~~2~~
L	~~1~~

$\sigma = 30\% \Leftrightarrow$ support ≥ 4

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB		CE	
AC		CF	
AD		CG	
AE		CH	
AF		DE	
AG		DF	
AH		DG	
BC		DH	
BD		EF	
BE		EG	
BF		EH	
BG		FG	
BH		FH	
CD		GH	

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB		CE	
AC		CF	
AD		CG	
AE		CH	
AF		DE	
AG		DF	
AH		DG	
BC		DH	
BD		EF	
BE	1	EG	1
BF		EH	1
BG	1	FG	
BH	1	FH	
CD		GH	1

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	1	CE	1
AC	1	CF	
AD		CG	1
AE	1	CH	1
AF		DE	
AG	1	DF	
AH	1	DG	
BC	1	DH	
BD		EF	
BE	2	EG	2
BF		EH	2
BG	2	FG	
BH	2	FH	
CD		GH	2

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	2	CE	2
AC	2	CF	1
AD		CG	1
AE	2	CH	2
AF	1	DE	
AG	1	DF	
AH	2	DG	
BC	2	DH	
BD		EF	1
BE	3	EG	2
BF	1	EH	3
BG	2	FG	
BH	3	FH	1
CD		GH	2

C2 and S2

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	2	CE	3
AC	2	CF	2
AD		CG	2
AE	2	CH	3
AF	1	DE	1
AG	1	DF	1
AH	2	DG	1
BC	3	DH	1
BD	1	EF	2
BE	4	EG	3
BF	2	EH	4
BG	3	FG	1
BH	4	FH	2
CD	1	GH	3

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	CH	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
BC	3	DH	1
BD	1	EF	2
BE	5	EG	3
BF	2	EH	5
BG	3	FG	1
BH	5	FH	2
CD	1	GH	3

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	CH	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
BC	3	DH	1
BD	1	EF	3
BE	6	EG	4
BF	3	EH	6
BG	4	FG	2
BH	6	FH	3
CD	1	GH	4

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	4	CE	3
AC	2	CF	2
AD	1	CG	2
AE	3	CH	3
AF	1	DE	1
AG	2	DF	1
AH	4	DG	2
BC	3	DH	2
BD	2	EF	3
BE	6	EG	4
BF	3	EH	6
BG	5	FG	2
BH	7	FH	3
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	1
AH	4	DG	3
BC	3	DH	2
BD	3	EF	3
BE	6	EG	4
BF	3	EH	6
BG	6	FG	2
BH	7	FH	3
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	2
AH	4	DG	4
BC	3	DH	2
BD	4	EF	3
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
BH	7	FH	3
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	4
AC	2	CF	3
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	2
AH	4	DG	4
BC	3	DH	2
BD	4	EF	4
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
BH	7	FH	3
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	5
AC	3	CF	4
AD	2	CG	2
AE	4	CH	4
AF	2	DE	1
AG	3	DF	2
AH	5	DG	4
BC	3	DH	2
BD	4	EF	5
BE	6	EG	4
BF	4	EH	7
BG	7	FG	3
BH	7	FH	4
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	6	CE	5
AC	3	CF	4
AD	2	CG	2
AE	5	CH	4
AF	2	DE	1
AG	4	DF	2
AH	5	DG	4
BC	3	DH	2
BD	4	EF	5
BE	7	EG	5
BF	4	EH	7
BG	8	FG	3
BH	7	FH	4
CD	1	GH	5

C2 and S2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

S2

AB	6	CE	5
AC	~~3~~	CF	4
AD	~~2~~	CG	~~2~~
AE	5	CH	4
AF	~~2~~	DE	~~1~~
AG	4	DF	~~2~~
AH	5	DG	4
BC	~~3~~	DH	~~2~~
BD	4	EF	5
BE	7	EG	5
BF	4	EH	7
BG	8	FG	~~3~~
BH	7	FH	4
CD	~~1~~	GH	5

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	
ABG	
ABH	
AEG	
AEH	
AGH	

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	
BDE	
BDF	
BDG	
BDH	
BEF	

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	CEF
BDE	CEH
BDF	CFH
BDG	
BDH	
BEF	

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	CEF
BDE	CEH
BDF	CFH
BDG	EFG
BDH	EFH
BEF	EGH

S2 to C3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	~~BFG~~
AEG	BFH
AEH	BGH
AGH	CEF
~~BDE~~	CEH
~~BDF~~	CFH
BDG	~~EFG~~
~~BDH~~	EFH
BEF	EGH

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE		BEH	
ABG		BFH	
ABH		BGH	
AEG		CEF	
AEH		CEH	
AGH		CFH	
BDG		EFH	
BEF		EGH	
BEG			

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE		BEH	1
ABG		BFH	
ABH		BGH	1
AEG		CEF	
AEH		CEH	
AGH		CFH	
BDG		EFH	
BEF		EGH	1
BEG	1		

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	1	BEH	2
ABG	1	BFH	
ABH	1	BGH	2
AEG	1	CEF	
AEH	1	CEH	1
AGH	1	CFH	
BDG		EFH	
BEF		EGH	2
BEG	2		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	2	BEH	3
ABG	1	BFH	1
ABH	2	BGH	2
AEG	1	CEF	1
AEH	2	CEH	2
AGH	1	CFH	1
BDG		EFH	1
BEF	1	EGH	2
BEG	2		

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	2	BEH	4
ABG	1	BFH	2
ABH	2	BGH	3
AEG	1	CEF	2
AEH	2	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	5
ABG	1	BFH	2
ABH	3	BGH	3
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	1	BFH	3
ABH	3	BGH	4
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	2	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	2	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	3	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	3
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	4
AEH	4	CEH	4
AGH	2	CFH	3
BDG	4	EFH	4
BEF	3	EGH	4
BEG	4		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	4	BEH	6
ABG	4	BFH	3
ABH	4	BGH	5
AEG	2	CEF	4
AEH	4	CEH	4
AGH	2	CFH	3
BDG	4	EFH	4
BEF	3	EGH	4
BEG	5		

C3 and S3

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S3

ABE	4	BEH	6
ABG	4	BFH	~~3~~
ABH	4	BGH	5
AEG	~~2~~	CEF	4
AEH	4	CEH	4
AGH	~~2~~	CFH	~~3~~
BDG	4	EFH	4
BEF	~~3~~	EGH	4
BEG	5		

S3 to C4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S3

ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

→

C4

S3 to C4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S3

ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4

ABEG
ABEH
ABGH

S3 to C4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S3

ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4

ABEG
ABEH
ABGH
BEGH

S3 to C4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S3

ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4

ABEG
ABEH
ABGH
BEGH
CEFH

S3 to C4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S3

ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

→

C4

~~ABEG~~
ABEH
~~ABGH~~
BEGH
~~CEFH~~

AEG, AGH and CFH not *frequent*!

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	
BEGH	

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	
BEGH	1

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	1
BEGH	2

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	2
BEGH	2

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	2
BEGH	3

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	3
BEGH	3

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	3
BEGH	4

C4 and S4

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

	1	B E G H
	2	A B C E G H
	3	A B C E F H
	4	B C D E F G H L
	5	A B E K H
	6	B E F G H I K
→	7	A B D G H
→	8	A B D G
→	9	B D F G
→	10	C E F
→	11	A C E F H
→	12	A B E G

C4

ABEH	3
BEGH	4

C4 and S4

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S4

ABEH	~~3~~
BEGH	4

Only one frequent 4-Itemset remaining.

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Outline

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

In Theory...

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- naïve algorithm: count frequency of all k-itemsets, for all $\binom{I}{k}$ k-itemsets, for all k
- number of possible itemsets $0 \leq k \leq |I|$?
- Apriori: one database scan for all frequent k-itemset *candidates* of a given k
- reduction of number of candidates by the anti-monotonicity principle of frequency: generate only candidates that have a chance to be frequent (join of frequent $(k-1)$-itemsets and pruning)

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Experiments

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Definition: Association Rule

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in $\mathcal{D}$ is the support of the union of its components: $s(X \Rightarrow Y) = s(X \cup Y)$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Problem 2: Association Rule Mining

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in $\mathcal{D}$ with a support of at least σ and a confidence of at least c, i.e.:

$$\{X \Rightarrow Y | s(X \Rightarrow Y) \geq \sigma \wedge \mathrm{conf}(X \Rightarrow Y) \geq c\}$$

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Find Association Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

This is part of the Apriori algorithm [Srikant and Agrawal, 1996].

for frequent itemset X:

- for each $Y \subset X$, $Y \neq \emptyset$, build the rule $Y \Rightarrow (X \setminus Y)$
- $\mathrm{conf}(Y \Rightarrow (X \setminus Y)) = \frac{s(X)}{s(Y)}$
- delete rules with confidence below a given threshold c

Note that:

For all involved itemsets (X, Y, $(X - Y)$), we have the support from the solution of problem 1 (stored or reconstructable from closed frequent itemsets). Thus we don't need a single database scan here.

Monotonicity of Confidence

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 2.1

Given:

- *itemset* X
- $Y \subset X, Y \neq \emptyset$

If $\mathrm{conf}(Y \Rightarrow (X \setminus Y)) < c$, *then* $\forall Y' \subset Y$*:*

$$\mathrm{conf}(Y' \Rightarrow (X \setminus Y')) < c.$$

This property allows the construction of all association rules satisfying some confidence threshold from all frequent itemsets with a procedure similar to the Apriori construction of frequent itemsets, but without database scan. [Srikant and Agrawal, 1996]

Pruning of Association Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Adapted from Tan et al. [2006], Fig. 6.15.

Example: Association Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Association rules for BEGH (see slide 88), confidence $\geq 60\%$, support 4:

Antecedent	Consequent	Support	Confidence	Rule
BEGH	$\emptyset$	4	1.000	
BEG	H	5	$4/5 =$ **0.800**	BEG $\Rightarrow$ H
BE	GH	7	$4/7 \approx$ 0.571	-
BG	EH	8	$4/8 =$ 0.500	-
EG	BH	5	$4/5 =$ **0.800**	EG $\Rightarrow$ BH
BEH	G	6	$4/6 \approx$ **0.667**	BEH $\Rightarrow$ G
BH	EG	7	$4/7 \approx$ 0.571	-
EH	BG	7	$4/7 \approx$ 0.571	-
BGH	E	5	$4/5 =$ **0.800**	BGH $\Rightarrow$ E
GH	BE	5	$4/5 =$ **0.800**	GH $\Rightarrow$ BE
EGH	B	4	$4/4 =$ **1.000**	EGH $\Rightarrow$ B

Interpretation of Support and Confidence

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

support: measures the frequency of the item set

- rules with very low support may occur simply by chance
- rules with low support are uninteresting from a business perspective

confidence: measures the reliability of the rule

- $X \Rightarrow Y$ – the higher the confidence, the more likely Y is present in transactions that contain X
- estimate of the conditional probability of Y given X

Limitations of Support and Confidence as Measures

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

	coffee	no coffee	
tea	150	50	200
no tea	650	150	800
	800	200	1000

$\{tea\} \Rightarrow \{coffee\}$

support?

confidence?

$\{\} \Rightarrow \{coffee\}$

support?

confidence?

Conclusion?

(Discussed by Tan et al. [2006], page 372f., example 6.3.)

Interestingness

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Other measures to assess the interestingness of a rule include:

Lift: $\mathit{Lift}(A \Rightarrow B) = \frac{\mathrm{conf}(A \Rightarrow B)}{f(B)}$

Jaccard: $\mathit{Jaccard}(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)+s(B)-s(A \cup B)}$

conviction: $\mathit{conviction}(A \Rightarrow B) = \frac{1-f(B)}{1-\mathrm{conf}(A \Rightarrow B)}$

Advanced reading: Vreeken and Tatti [2014], Tan et al. [2006] (Ch. 6), Witten et al. [2011] (Ch. 6.3)

Outline

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
- Sets and Relations
- Item Sets
- Definitions and Problem-Statements
- Monotonicity-Property of Frequency
- The Apriori Algorithm
- Example
- Efficiency
- Association Rule Mining
- Summary

Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outlook: Other Algorithms

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Frequent pattern mining is a large research field, for an overview see the collection of topics edited by Aggarwal and Han [2014].
- Another important algorithm for frequent pattern mining is FP-Growth [Han et al., 2000].
- Frequent patterns have been defined in other application scenarios such as, e.g., graphs, spatiotemporal data, sequential data (for example protein sequences, as in the work of Birzele and Kramer [2006]).
- The principle of anti-monotonicity for pruning has been applied in many other application areas (e.g., in subspace clustering [Zimek et al., 2014]).

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and Problem-Statements
Monotonicity-Property of Frequency
The Apriori Algorithm
Example
Efficiency
Association Rule Mining
Summary
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *mathematical basics on sets, relations, and orders*
- *the problem of frequent pattern mining*
- *the problem of association rule mining*
- *the "Apriori principle" (the anti-monotonicity property of frequent itemsets)*
- *the Apriori algorithm for mining of frequent itemsets and association rules*
- *pros and cons of support and confidence*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Zaki and Meira Jr. [2014], Chapters 2+3*
- *Tan et al. [2006], Chapters 2+3*
- *Han et al. [2011], Chapter 2*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The KDD process model

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

KDD process model (cf. Fayyad et al. [1996])

- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - discretize numeric attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Deriving Features from Complex Objects

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: CAD-objects

possible features:

- width
- height
- curvature parameters a, b, c

Deriving Features from Complex Objects

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Transformation:

object space $\longrightarrow$ feature space

- features are combined to feature vectors
- often high-dimensional feature spaces (here only 5-dim.)
- statistical context: features are called ‘variables’

Scale Characteristics of Features

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- nominal (categories):
  - It is possible to tell whether two values are equal or not – but no direction (better, more,. . .) or meaningful distance.
  - Examples: sex, eye color, healthy/sick, amino acids
- ordinal
  - there is an order relation (e.g., better/worse), but no uniform distance
  - Examples: grade, quality label, age class (e.g., 20-29, 30-39,. . .), color (?)
- metric
  - differences and relations between values are meaningful, values can be discrete or continuous
  - Examples: weight, selling counts, age

Aggregations of Features

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For a feature X, we have a sample $x_1, x_2, \ldots, x_n$.

- absolute frequency: For each value a, we have $f(a)$ as the number of occurrences in the sample.
- relative frequency: $p(a) = f(a)/n$
- mode: the value with largest frequency

for at least ordinal features:

- median: the central element in the sample ordered by value

for metric features:

- arithmetic mean: $\mu = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$

Skewness

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Aggregations of Feature Vectors

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For a given sample $\mathcal{D}$, we have:

- centroid: $\mu_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \cdot \sum_{o \in \mathcal{D}} o$

  In a general metric space (that is, not a vector space), where we only have pairwise distances, it might not be possible to compute a centroid.

- medoid: $\arg\min_{o \in \mathcal{D}} \mathrm{dist}(o, \mu_{\mathcal{D}})$

  In a general metric space, the medoid is the object with the smallest average distance to all other objects.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Feature Space and Distance Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A feature space is a domain with a distance function:

$$F = (\mathrm{dom}, \mathrm{dist})$$

- dom is a sorted set of features
- $\mathrm{dist} : \mathrm{dom} \times \mathrm{dom} \rightarrow \mathbb{R}_0^+$ is a total (distance) function with the following properties:
  - strictness:

$$\forall p, q \in \mathrm{dom}, p \neq q : \mathrm{dist}(p, q) > 0$$

  - reflexivity:

$$\forall o \in \mathrm{dom} : \mathrm{dist}(o, o) = 0$$

  - symmetry:

$$\forall p, q \in \mathrm{dom} : \mathrm{dist}(p, q) = \mathrm{dist}(q, p)$$

Metric Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $M = (\mathrm{dom}, \mathrm{dist})$ is a metric space, if the following properties are given:
  - M is a feature space
  - the triangle inequality holds:

$$\forall o, p, q \in \mathrm{dom} : \mathrm{dist}(o, p) \leq \mathrm{dist}(o, q) + \mathrm{dist}(q, p)$$

- most common example: Euclidean *vector* space:
  - $\mathrm{dom} = \mathbb{R}^d$ (d: number of features – dimensionality)
  - $\mathrm{dist} = (x, y) \mapsto \|x - y\|_2$

Spaces and Distance Functions

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Common distance measure for (Euclidean) feature vectors: $L_{\mathcal{P}}$-norm

$$\mathrm{dist}_{\mathcal{P}}(p,q) = \left(|p_1 - q_1|^{\mathcal{P}} + |p_2 - q_2|^{\mathcal{P}} + \ldots + |p_n - q_n|^{\mathcal{P}}\right)^{\frac{1}{\mathcal{P}}}$$

Euclidean norm (L_2):

Manhattan norm (L_1):

Maximum norm (L_∞, also: $L_{\max}$, supremum dist., Chebyshev dist.)

Spaces and Distance Functions

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

weighted Euclidean norm:

$$\mathrm{dist}(p,q) = \left(w_1|p_1 - q_1|^2 + w_2|p_2 - q_2|^2 + \ldots + w_n|p_n - q_n|^2\right)^{\frac{1}{2}}$$

quadratic form*:

$$\mathrm{dist}(p,q) = \left((p-q)M(p-q)^{\mathsf{T}}\right)^{\frac{1}{2}}$$

*note that we assume vectors to be row vectors here

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Categories of Feature Descriptors for Images

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- distribution of colors
- texture
- shapes (contoures)

Color Histogram

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- a histogram represents the distribution of colors over the pixels of an image
- definition of an color histogram:
  - choose a color space (RGB, HSV, HLS, ...)
  - choose number of representants (sample points) in the color space
  - possibly normalization (to account for different image sizes)

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

original images in full RGB space ($256^3 = 16,777,216$)

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

2^3

3^3

4^3

16^3

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

2^3

3^3

4^3

16^3

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

2^3

3^3

4^3

16^3

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

2^3

3^3

4^3

16^3

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The histogram for each image is essentially a visualization of a vector:

$(0.77, 0, 0, 0, 0.08, 0, 0.15, 0)$ $(0.8, 0, 0, 0, 0.11, 0, 0.09, 0)$

$(0.9, 0, 0, 0, 0.05, 0, 0.05, 0)$ $(0.955, 0, 0, 0, 0.045, 0, 0, 0)$

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Impact of Number of Representants

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distances for Color Histograms

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Euclidean distance for images P and Q using the color histograms h_P and h_Q:

$$\mathrm{dist}(P,Q) = \sqrt{(h_P - h_Q)\cdot(h_P - h_Q)^\mathsf{T}}$$

$$\mathrm{dist}(\text{RED}, \text{PINK}) = \sqrt{2}$$
$$\mathrm{dist}(\text{RED}, \text{BLUE}) = \sqrt{2}$$
$$\mathrm{dist}(\text{BLUE}, \text{PINK}) = \sqrt{2}$$

A 'psychologic' distance would consider that red is (in our perception) more similar to pink than to blue.

Distances for Color Histograms

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Quadratic form with 'psychological' similarity matrix

$$A = \begin{bmatrix} 1 & a_{12} & \dots & \\ a_{21} & 1 & \dots & \\ \vdots & & \ddots & \vdots \\ & & \dots & 1 \end{bmatrix}$$

where a_{ij} ($\stackrel{?}{=} a_{ji}$) describe the subjective similarity of the features i and j in the color histogram:

$$\mathrm{dist}_A(P,Q) = \sqrt{(h_P - h_Q) \cdot A \cdot (h_P - h_Q)^\mathsf{T}}$$

$$A' = \begin{bmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathrm{dist}(\text{RED}, \text{PINK}) = \sqrt{0.2}$$

$$\mathrm{dist}(\text{RED}, \text{BLUE}) = \sqrt{2}$$

$$\mathrm{dist}(\text{BLUE}, \text{PINK}) = \sqrt{2}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Texts as "Bag of Words"

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- texts can be seen as sets of words or vectors of terms
- a term can be
  - a single word
  - a phrase or part of a sentence
- typical feature extraction from texts: transform a text (document) into a vector of term frequencies:

$$d \mapsto (f_{t_1 d}, f_{t_2 d}, \ldots, f_{t_n d})$$

where $f_{t_i d}$ is the frequency of term t_i in document d.

The aim of machine learning is to build computer systems that can adapt to their environments and learn from experience. Learning techniques and methods from this field are successfully applied to a variety of learning tasks in a broad range of areas, including, for example, spam recognition, text classification, gene discovery, financial forecasting.

$$\mapsto \begin{pmatrix} \vdots \\ \text{machine: } 1 \\ \text{learn: } 4 \\ \text{to: } 3 \\ \vdots \end{pmatrix}$$

Problems with Term Frequencies

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

1. Many words are totally pointless for distinguishing texts (the, a, it, that, . . .).
2. The same word can appear differently (learn, learning; go, went).
3. The feature space typically becomes very high dimensional (often $n > 10000$).
4. Not all terms are equally useful.
5. Most words from a dictionary don't occur at all in any of the compared documents ($f = 0$).

additional linguistic problems:

- different words with the same (or closely similar) meaning ("buy", "purchase")
- words with very different meaning ("mouse")

Problem 1

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Many words are totally pointless for distinguishing texts (the, a, it, that, . . .).
- Solution: eliminate these words before computing the term frequency vector.
- Such words are called stopwords, list of stopwords for many languages exist.

Problem 2

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The same word can appear differently (learn, learning; go, went).
- Solution: stemming. Any word is mapped to its stem (base or root form).
- For English texts, algorithmic stemming is possible (Porter's stemming algorithm, see `http://tartarus.org/~martin/PorterStemmer/index.html`).
- For other languages, dictionaries are required.

Problem 3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The feature space typically becomes very high dimensional (often $d > 10000$).
- Solution: selection of the most important terms (feature selection).
- Example: intermediate document frequency
  - very frequent terms appear in almost all documents
  - very rare terms appear in almost no document

  In both cases, these terms are not helpful to distinguish most of the documents.

Problem 3

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Approach:

1. document frequency for all terms: $DF(t_i) = \frac{|\{d|t_i \in d\}|}{|\mathcal{D}|}$
2. sort terms according to $DF(t_i)$

Rank	Term	DF
1.	t_{23}	0.82
2.	t_{17}	0.65
3.	t_{14}	0.52
4.	...	...

3. sort terms according to score$(t_i) = DF(t_i) \cdot$ rank(t_i) examples:

$$\text{score}(t_{23}) = 0.82 \cdot 1$$
$$\text{score}(t_{17}) = 0.65 \cdot 2$$

4. chose the k terms with largest score

Problem 4

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Not all terms are equally useful.
- Solution: TF-IDF (term frequency – inverse document frequency)
  - higher weights for rare words
  - higher weights for terms that are more frequent than others in some document
- TF is the relative term frequency in some document d: $TF(t,d) = \frac{n(t,d)}{\sum_{t_i \in d} n(t_i,d)}$
- IDF is the inverse document frequency of t for all documents: $IDF(t) = \frac{|\mathcal{D}|}{|\{d | d \in \mathcal{D} \wedge t \in d\}|}$
- feature vector for document $d = \begin{pmatrix} TF(t_1,d) \cdot IDF(t_1) \\ TF(t_2,d) \cdot IDF(t_2) \\ \vdots \\ TF(t_k,d) \cdot IDF(t_k) \end{pmatrix}$

Problem 5

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Most words from a dictionary don't occur at all in any of the compared documents ($f = 0$).
- As a consequence, the Euclidean distance will be very similar for many documents.
- Solution: use different distance measures
- Jaccard coefficient: documents as sets of terms:

$$\mathrm{dist}_{\text{Jaccard}}(D_1, D_2) = 1 - \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

- cosine coefficient (possibly for TF-IDF vectors):

$$\mathrm{dist}_{\text{cosine}}(D_1, D_2) = 1 - \frac{\langle D_1, D_2 \rangle}{\|D_1\| \cdot \|D_2\|}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
- Features: Descriptors for Objects
- Feature Space
- Features for Images
- Features for Texts
- Summary

Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Your Choice of a Distance Measure

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

There are hundreds of distance functions [Deza and Deza, 2009].

- For time series: DTW, EDR, ERP, LCSS, ...
- For texts: Cosine and normalizations
- For sets – based on intersection, union, ... (Jaccard)
- For clusters (single-link, average-link, etc.)
- For histograms: histogram intersection, "Earth movers distance", quadratic forms with color similarity
- With normalization: Canberra, ...
- Quadratic forms: $d(x,y) := \sqrt{(x-y)M(x-y)^\top}$ for some positive (usually symmetric) definite matrix M.

Note that:

Choosing the appropriate distance function can be seen as a part of "preprocessing".

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Features: Descriptors for Objects
Feature Space
Features for Images
Features for Texts
Summary
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *features and feature spaces*
- *categories of features (categorical/nominal, ordinal, metric)*
- *basic univariate feature descriptors (frequency (relative/absolute), mode, median, mean)*
- *distances (L_p-norms, weighted, quadratic form)*
- *feature (vector) descriptors for texts and for images*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Tan et al. [2006], Chapter 8.1+8.2*
- *Han et al. [2011], Chapter 10.1+10.2*
- *Zaki and Meira Jr. [2014], Chapter 13.1*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Purpose of Clustering

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- identify a finite number of categories (classes, groups: “clusters”) in a given dataset
- *similar* objects shall be grouped in the same cluster, *dissimilar* objects in different clusters
- “similarity” is highly subjective, depending on the application scenario

Clustering is Unsupervised

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Clustering is unsupervised, that is, we do not have any external knowledge to guide or to supervise the clustering process:

- we cannot learn rules to sort points into clusters
- we do not know how many clusters there are
- we do not know how the clusters are characterized
- there is no unique criterion to judge on the quality of a derived clustering solution (evaluation)

Challenges

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Assume we are given a hypothetical quality function q to decide for a given partition of a set of n points whether or not this partition constitutes a (good) clustering.

- naïve method: test q on all possible clusterings with k partitions (clusters) ($2 \leq k \leq ?$)
- problems:
  - there are $\mathcal{O}(k^n)$ many partitions in k clusters
  - and we don't have this function q, actually

Therefore, we need heuristic solutions for both problems:

- *efficient* search for solutions
- efficient and effective modelling of q

There are many such heuristic solutions around, that is, we have a plethora of clustering algorithms in the literature.

Categories of Clustering Approaches

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- partitioning
  - model: clusters are compact sets of points
  - parameter: (usually) number k of clusters, distance measure
  - looks for a flat partitioning into k cluster with maximal compactness

Categories of Clustering Approaches

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- density-based
  - model: clusters are areas of high density, separated by areas of low density
  - parameter: minimal density in some cluster, distance measure
  - looks for a flat partitioning into clusters exceeding some minimal density

Categories of Clustering Approaches

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- hierarchical
  - model: compactness, density,...
  - parameter: distance measure for points and for clusters
  - looks for a hierarchy of clusters (e.g., given as a tree), joins the most similar clusters at a given level of the hierarchy
  - flat clusters can be derived by cutting the tree on some level

Categories of Clustering Approaches

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Alternative categorizations could sort approaches according to the algorithmic approach or to the application scenario of specialized approaches:

- many variants of techniques, such as fuzzy clustering, graph-theoretic algorithms, neuronal nets
- specializations to special data characteristics (time series, graphs, high-dimensional data, stream data, uncertain data etc.)

examples: Aggarwal and Reddy [2013]

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Objective

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Partitional clustering partitions a dataset into k clusters, typically minimizing some cost function (compactness criterion).

Central assumptions for approaches in this family are typically:

- number k of clusters known
- clusters are characterized by their compactness
- compactness measured by some distance function (e.g., distance of all objects in a cluster from some cluster representative is minimal)
- criterion of compactness typically leads to convex or even spherically shaped clusters

Basic Strategy

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A complete search for the globally optimal solution is intractable, thus:

- locally optimizing algorithms
  - choose k initial representatives
  - assign each object to its closest representative
  - optimize the representatives iteratively
- types of cluster representatives
  - mean (centroid) – central object is constructed
  - some cluster object (e.g., medoid) – central object is chosen
  - Gaussian distribution model for a cluster – expectation maximization

Construction of Central Points: Example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Construction of Central Points: Basics

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- objects are points $x = (x_1, \ldots, x_d)$ in Euclidean vector space $\mathbb{R}^d$, dist = Euclidean distance (L_2)
- centroid μ_C: mean vector of all points in cluster C

$$\mu_{C_i} = \frac{1}{|C_i|} \cdot \sum_{o \in C_i} o$$

Construction of Central Points: Basics

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- objects are points $x = (x_1, \ldots, x_d)$ in Euclidean vector space $\mathbb{R}^d$, dist = Euclidean distance (L_2)
- centroid μ_C: mean vector of all points in cluster C

$$\mu_{C_i} = \frac{1}{|C_i|} \cdot \sum_{o \in C_i} o$$

Construction of Central Points: Basics

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- measure of compactness for a cluster C:

$$TD^2(C) = \sum_{p \in C} \mathrm{dist}(p, \mu_C)^2$$

(a.k.a. SSQ: sum of squares or SSE: sum of squared error)

- measure of compactness for a clustering

$$TD^2(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} TD^2(C_i)$$

Construction of Central Points: Basics

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- measure of compactness for a cluster C:

$$TD^2(C) = \sum_{p \in C} \mathrm{dist}(p, \mu_C)^2$$

(a.k.a. SSQ: sum of squares or SSE: sum of squared error)

- measure of compactness for a clustering

$$TD^2(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} TD^2(C_i)$$

Algorithm: Clustering by Minimization of Variance [Forgy, 1965, Lloyd, 1982]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 4.1 (Clustering by Minimization of Variance)

1. *start with k (e.g., randomly selected) points as cluster representatives (or a random partition in k "clusters")*
2. *repeat:*
   - 2.1 *assign each point to the closest representative*
   - 2.2 *compute new representatives based on the given partitions (centroid of the assigned points)*

   until there is no change in assignment

Model

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The assignment to the closest representative corresponds to a Voronoi parcelling

Model

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Voronoi parcelling $\neq$ convex hull!

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Convergence: Visualization of Evolving Voronoi Parcelling

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k-means

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k-means [MacQueen, 1967] is the best known variant of the basic algorithm:

- a centroid is immediately updated when some point changes its assignment
- k-means has very similar properties, but the result now depends on the order of data points in the input file

Note that:

The name "k-means" is often used indifferently for any variant of the basic algorithm, in particular also for the algorithm of Forgy [1965], Lloyd [1982].

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (6.0, 4.3)$$

$$\mu \approx (5.0, 6.4)$$

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign points

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (5.0, 2.7)$$

$$\mu \approx (5.6, 7.4)$$

k-means Clustering – Lloyd/Forgy Algorithm

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign points

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (4.0, 3.25)$$

$$\mu \approx (6.75, 8.0)$$

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign points

k-means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign points
no change
convergence!

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Centroids (e.g.: from previous iteration):

$$\mu \approx (6.0, 4.3)$$

$$\mu \approx (5.0, 6.4)$$

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign first point

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign second point

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (5.0, 4.0)$$

$$\mu \approx (5.75, 7.25)$$

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign third point

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$\mu \approx (4.6, 4.0)$

$\mu \approx (6.7, 8.3)$

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign fourth point

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assing fifth point

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (4.0, 3.25)$$

$$\mu \approx (6.75, 8.0)$$

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign more points

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign more points
possibly more iterations

k-means Clustering – MacQueen Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign more points
possibly more iterations
convergence

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Centroids (e.g.: from previous iteration):

$$\mu \approx (6.0, 4.3)$$

$$\mu \approx (5.0, 6.4)$$

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign first point

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign second point

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign third point

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign fourth point

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$\mu \approx (4.0, 8.5)$

$\mu \approx (4.3, 6.2)$

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

assign fifth point

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$$\mu \approx (10.0, 1.0)$$

$$\mu \approx (4.7, 6.3)$$

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reasign more points

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reasign more points
possibly more iterations

k-means Clustering – MacQueen Algorithm

Alternative Run – Different Order

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reasign more points
possibly more iterations
convergence

k-means Clustering – Quality

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$SSQ(\mu_1, p_1) = |4-10|^2 + |3.25-1|^2 = 36 + 5\tfrac{1}{16} = 41\tfrac{1}{16}$$
$$SSQ(\mu_1, p_2) = |4-2|^2 + |3.25-3|^2 = 4 + \tfrac{1}{16} = 4\tfrac{1}{16}$$
$$SSQ(\mu_1, p_3) = |4-3|^2 + |3.25-4|^2 = 1 + \tfrac{9}{16} = 1\tfrac{9}{16}$$
$$SSQ(\mu_1, p_4) = |4-1|^2 + |3.25-5|^2 = 9 + 3\tfrac{1}{16} = 12\tfrac{1}{16}$$
$$TD^2(C_1) = 58\tfrac{3}{4}$$

$$SSQ(\mu_2, p_5) = |6.75-7|^2 + |8-7|^2 = \tfrac{1}{16} + 1 = 1\tfrac{1}{16}$$
$$SSQ(\mu_2, p_6) = |6.75-6|^2 + |8-8|^2 = \tfrac{9}{16} + 0 = \tfrac{9}{16}$$
$$SSQ(\mu_2, p_7) = |6.75-7|^2 + |8-8|^2 = \tfrac{1}{16} + 0 = \tfrac{1}{16}$$
$$SSQ(\mu_2, p_8) = |6.75-7|^2 + |8-9|^2 = \tfrac{1}{16} + 1 = 1\tfrac{1}{16}$$
$$TD^2(C_2) = 2\tfrac{3}{4}$$

First solution: $TD^2 = 61\frac{1}{2}$

Note: $SSQ(\mu, p) = \textit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

k-means Clustering – Quality

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$SSQ(\mu_1, p_1) = |10 - 10|^2 + |1 - 1|^2 = 0$
$TD^2(C_1) = 0$

$SSQ(\mu_2, p_2) \approx |4.7 - 2|^2 + |6.3 - 3|^2 \approx 18.2$
$SSQ(\mu_2, p_3) \approx |4.7 - 3|^2 + |6.3 - 4|^2 \approx 8.2$
$SSQ(\mu_2, p_4) \approx |4.7 - 1|^2 + |6.3 - 5|^2 \approx 15.4$
$SSQ(\mu_2, p_5) \approx |4.7 - 7|^2 + |6.3 - 7|^2 \approx 5.7$
$SSQ(\mu_2, p_6) \approx |4.7 - 6|^2 + |6.3 - 8|^2 \approx 4.6$
$SSQ(\mu_2, p_7) \approx |4.7 - 7|^2 + |6.3 - 8|^2 \approx 8.2$
$SSQ(\mu_2, p_7) \approx |4.7 - 7|^2 + |6.3 - 9|^2 \approx 12.6$
$TD^2(C_2) \approx 72.86$

First solution: $TD^2 = 61\frac{1}{2}$
Second solution: $TD^2 \approx 72.68$

Note: $SSQ(\mu, p) = \textit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

k-means Clustering – Quality

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$SSQ(\mu_1, p_2) = |2-2|^2 + |4-3|^2 = 0 + 1 = 1$
$SSQ(\mu_1, p_3) = |2-3|^2 + |4-4|^2 = 1 + 0 = 1$
$SSQ(\mu_1, p_4) = |2-1|^2 + |4-5|^2 = 1 + 1 = 2$
$TD^2(C_1) = 4$

$SSQ(\mu_2, p_1) = |7.4-10|^2 + |6.6-1|^2 = 6\frac{19}{25} + 31\frac{9}{25} = 38\frac{3}{25}$
$SSQ(\mu_2, p_5) = |7.4-7|^2 + |6.6-7|^2 = \frac{4}{25} + \frac{4}{25} = \frac{8}{25}$
$SSQ(\mu_2, p_6) = |7.4-6|^2 + |6.6-8|^2 = 1\frac{24}{25} + 1\frac{24}{25} = 3\frac{23}{25}$
$SSQ(\mu_2, p_7) = |7.4-7|^2 + |6.6-8|^2 = \frac{4}{25} + 1\frac{24}{25} = 2\frac{3}{25}$
$SSQ(\mu_2, p_8) = |7.4-7|^2 + |6.6-9|^2 = \frac{4}{25} + 5\frac{19}{25} = 5\frac{23}{25}$
$TD^2(C_2) = 50\frac{2}{5}$

First solution: $TD^2 = 61\frac{1}{2}$
Second solution: $TD^2 \approx 72.68$
Optimal solution: $TD^2 = 54\frac{2}{5}$

Note: $SSQ(\mu, p) = \textit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

pros

- efficient: $\mathcal{O}(k \cdot n)$ per iteration, number of iterations is usually in the order of 10.
- easy to implement, thus very popular

cons

- k-means converges towards a *local* minimum
- k-means (MacQueen-variant) is order-dependent
- deteriorates with noise and outliers (all points are used to compute centroids)
- clusters need to be convex and of (more or less) equal extension
- number k of clusters is hard to determine
- strong dependency on initial partition (in result quality as well as runtime)

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Selection of Representative Points: k-medoids

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

works in a general metric space:

- requires only distance measure dist for pairs of objects
- medoid as representative: a central object of the cluster
- measure of compactness for a cluster C with medoid m_C:

$$TD(C) = \sum_{p \in C} \mathrm{dist}(p, m_C)$$

- measure of compactness for a clustering

$$TD(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} TD(C_i)$$

- The medoid of a cluster is the point among the cluster members that minimizes TD.

k-modes [Huang, 1997]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For categorical data: use the mode as an analogue to the centroid:

- in numerical attributes: centroid $\bar{x}$ of C minimizes $TD^2(C, \bar{x}) = \sum_{p \in C} \mathrm{dist}(p, \bar{x})^2$
- categorical attributes: mode m of C minimizes $TD(C, m) = \sum_{p \in C} \mathrm{dist}(p, m)$, where m is not necessarily an element of C
- distance function for categorical attributes, e.g.:

$$\mathrm{dist}(x, y) = \sum_{i=1}^{d} \delta(x_i, y_i), \text{ where } \delta(x_i, y_i) = \begin{cases} 0 & \text{, if } x_i = y_i \\ 1 & \text{, else} \end{cases}$$

Find the Mode

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $TD(C, m) = \sum_{p \in C} \mathrm{dist}(p, m)$ is minimized, iff for $m = (m_1, \ldots, m_d)$ and all attributes $A_i, i = 1, \ldots, d$ holds: There is no attribute value more frequent than m_i in A_i.
- the mode is not necessarily unique
- example: $\mathcal{D} = \{(a, b), (a, c), (c, b), (b, c)\}$
- possible modes:

Find the Mode

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $TD(C, m) = \sum_{p \in C} \mathrm{dist}(p, m)$ is minimized, iff for $m = (m_1, \ldots, m_d)$ and all attributes $A_i, i = 1, \ldots, d$ holds: There is no attribute value more frequent than m_i in A_i.
- the mode is not necessarily unique
- example: $\mathcal{D} = \{(a, b), (a, c), (c, b), (b, c)\}$
- possible modes: $(a, b), (a, c)$

Algorithm k-modes

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- initialization: randomly selected objects as initial modes
- use mode instead of centroid
- use categorical distance measure instead of squared Euclidean distance

Then follow Algorithm 4.1.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Choice of Initial Model [Fayyad et al., 1998]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

clustering on a small sample usually delivers good initial clusters, but some samples are deviating too much from the overall data distribution

- draw independently m samples
- clustering of each sample results in m estimates of the k cluster centers
- run clustering algorithm on $\mathcal{D}$ m-times (with the m estimated start configurations)
- choose the clustering result with the best value for some quality criterion

Example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$k = 3$ Gaussians

centers of $m = 4$ samples
X: true centers

k-means++ Initialization [Arthur and Vassilvitskii, 2007]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 4.2 (k-means++ [Arthur and Vassilvitskii, 2007])

1. *Choose a first centroid randomly.*
2. *Compute for each point the distance to the closest of the already existing centroids.*
3. *Choose a new centroid from the points with a probability proportional to the squared distance.*
4. *Repeat 2 and 3 until k centroids have been chosen.*
5. *Continue with step 2 of Algorithm 4.1.*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Choice of k

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Method:

- do a clustering for $k = 2, 3, 4, \ldots, n-1$ (or some other reasonable upper limit, e.g., $n/2$, $\sqrt{n}$)
- choose the best of the results

What is "the best"?

- quality measure needs to be independent of k
- k-means, k-medoid etc.: TD^2 and TD improve (i.e., become smaller) with increasing k

Evaluation of Clusterings

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For the evaluation of clusterings we are mainly interested in two aspects:

cohesion: how tightly are members of one cluster connected, how compact are clusters?

separation: how well are different clusters separated?

Validity of Clusterings in General

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A validity measure for clusterings is typically some combination of cohesion and separation.
- Weights w_i could consider, e.g., the cluster size.

For a clustering C with k clusters $C_1, \ldots, C_k$, e.g.:

$$\text{overall_validity}(C) = \sum_{i=1}^{k} w_i \cdot \text{validity}(C_i)$$

Validity based on some measure of proximity (similarity-, distance-function):

$$\text{cohesion}(C_i) = \sum_{(x,y) \in C_i \times C_i} \text{proximity}(x, y)$$

$$\text{separation}(C_i, C_j) = \sum_{(x,y) \in C_i \times C_j} \text{proximity}(x, y)$$

Silhouette [Rousseeuw, 1987]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- let $a(o)$ be the average distance between o and all other members of the same cluster
- let $b(o)$ be the average distance between o and all members of another cluster that minimizes this average
- the silhouette of o is given by

$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}$$

- it holds that $-1 \leq s(o) \leq 1$
- $s(o) \approx -1, 0, 1$: bad, indifferent, good assignment of o

Simplification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

measure proximity w.r.t. the prototypes c_i of clusters C_i:

$$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i)$$

$$\text{separation}(C_i, C_j) = \sum_{x) \in C_i} \text{proximity}(x, c_j)$$

Even more simplified:

$$\text{separation}(C_i, C_j) = \text{proximity}(c_i, c_j)$$

Simplified Silhouette Coefficient [Rousseeuw, 1987]: Points

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- let $a(o)$ be the distance between o and its "own" cluster representative
- let $b(o)$ be the distance between o and the closest "foreign" cluster representative
- the silhouette of o is given by

$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}$$

- it holds that $-1 \leq s(o) \leq 1$
- $s(o) \approx -1, 0, 1$: bad, indifferent, good assignment of o

Example: Simplified Silhouette of Individual Points

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 8.29. Silhouette coefficients for points in ten clusters.

(Figure from Tan et al. [2006].)

Simplified Silhouette Coefficient [Rousseeuw, 1987]: Clustering

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- given the individual silhouette values $s(o)$
- the silhouette s_C of a clustering C is the average silhouette of all n points:

$$s_C = \frac{1}{n} \sum_{o \in \mathcal{D}} s(o)$$

- interpretation: $s_C > 0.7$: strong cluster structure, $s_C > 0.5$: reasonable cluster structure

Comparison: Silhouette – TD^2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 8.29. Silhouette coefficients for points in ten clusters.

Examples: Silhouette – TD^2

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

k=2
TD=6472,75
SC=0,506355

k=4
TD=3455,42
SC=0,552083

k=2 (best betw. 2 and 20)
TD=6725,05
SC=0,658093

k=3
TD=5117,11
SC=0,364401

k=9 (best betw. 2 and 20)
TD=2613,09
SC=0,395389

k=3
TD=5749,59
SC=0,429931

k=2 (best betw. 2 and 20)
TD=7314,49
SC=0,734972

Visual Evaluation: Silhouette Plot

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Example by Pedregosa et al. [2011].)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

Visual Evaluation: Silhouette Plot

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Example by Pedregosa et al. [2011].)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

Visual Evaluation: Silhouette Plot

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Example by Pedregosa et al. [2011].)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

Visual Evaluation: Silhouette Plot

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Example by Pedregosa et al. [2011].)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

Visual Evaluation: Silhouette Plot

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Example by Pedregosa et al. [2011].)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

A Problem for the k-Means Clustering – Lloyd/Forgy Algorithm

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign objects

A Problem for the k-Means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign objects

A Problem for the k-Means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:

$\mu = (2.75, 4.25)$

$\mu = (8.0, 6.0)$

$\mu = (12.0, 5.0)$

A Problem for the k-Means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

reassign objects

A Problem for the k-Means Clustering – Lloyd/Forgy Algorithm

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recompute centroids:
Oops.
Average of an empty set?

Strategies for Handling Empty Clusters

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

There is no golden (standard) solution. Possibilities are:

- Report one cluster less.
- Choose the point that is farthest away from any current centroid as a new centroid.
- Choose (randomly) a point from that cluster that currently exhibits the largest TD^2 as a new centroid.

Note that:

Each of these solutions has a certain impact on the sum of squared error (TD^2, the clustering objective function).
For which of these solutions would you argue?

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
- Introduction
- Basic Algorithms for Partitional Clustering
- Visualization: Algorithmic Differences
- Variants for Non-metric Spaces
- Initialization
- Model Selection and Evaluation
- Empty Clusters
- Summary

Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Opportunities for Further Research

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The silhouette is one of the most prominent, but just one among many internal clustering evaluation measures. Many measures are discussed by Vendramin et al. [2010].
- Although k-means is a time-honored, classic method, there is still a lot of research around it, for example to make it faster. Elkan [2003] proposes a speed-up based on properties of metric spaces.
- On the other hand, comparing runtimes of algorithms empirically is a bit tricky. Kriegel et al. [2017] discuss issues around that.

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Introduction
Basic Algorithms for Partitional Clustering
Visualization: Algorithmic Differences
Variants for Non-metric Spaces
Initialization
Model Selection and Evaluation
Empty Clusters
Summary
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *What is Clustering?*
- *Why do we need heuristic approaches to "solve" the clustering problem?*
- *Basic heuristic ideas for identifying "partitions" into k clusters*
  - *selection of representative points*
  - *optimization approaches for assignment of points to representatives:*
    - *minimization of variance [Forgy, 1965, Lloyd, 1982]*
    - *k-means [MacQueen, 1967]*
    - *k-medoids*
    - *k-modes*
  - *common ideas and differences between these approaches*
  - *evaluation, model selection, silhouette-coefficient*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Classification – Supervised Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- For some domain $\mathcal{D}$ and a set of classes $C = \{c_1, \ldots, c_k\}$, $k \geq 2$, each $o \in \mathcal{D}$ belongs uniquely to some $c \in C$, i.e., there is a function $f : \mathcal{D} \to C$ (see Slide 55).
- Given a set of objects $O = \{o_1, o_2, \ldots, o_n\} \subseteq \mathcal{D}$ and a mapping $(O \to C) \subset f$ (examples):
  We want to also map any object $o_m \in \mathcal{D} \setminus O$ to C.
- Supervised vs. unsupervised:
  - In clustering, we don't have any information on C.
  - In classification, we have examples (a training set) to guide (supervise) the learning process.

Classification – Supervised Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A classifier is trained on some training set $TR \subseteq O$ to learn the mapping function (a model or hypothesis) $h : \mathcal{D} \rightarrow C$.
- Ideally we have $\forall o \in TR : h(o) = f(o)$ (if not for all, we should have this at least for most examples o).
- After training, the classifier should also work on $\mathcal{D} \setminus TR$ and predict the correct class, i.e., $h \approx f$.
- f is called "the target function".

Assumption 5.1 (The inductive learning assumption)

Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other examples.

Training of a Classifier

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Application of a Classifier

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Intuition

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Data $\subset \mathcal{D} \times C$ are generated by some (natural / technical / social / ...) statistical process.
- The observed data are examples for the effect of the process.
- The challenge for the learning algorithm is to generate a model to explain the process.
- h is an approximation of f, a hypothesis to explain the data.
- In the ideal case, the hypothesis h is interpretable and helps to understand the data-generating process.
- Pragmatically, h might be useful for predictions although it might not be interpretable.

Classification Algorithms and their Hypothesis-Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Based on training data, a classifier typically provides a hypothesis that separates the examples belonging to different classes from each other.
- Each classification algorithm comes with (more or less strict) assumptions on how separation can be achieved or defined.

Classification Algorithms and their Hypothesis-Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- These assumptions (implicitly) define a hypothesis space $\mathcal{H}$, the space of hypotheses that could possibly be learned by the given algorithm.

Classification Algorithms and their Hypothesis-Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- These assumptions (implicitly) define a hypothesis space $\mathcal{H}$, the space of hypotheses that could possibly be learned by the given algorithm.

Classification Algorithms and their Hypothesis-Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- These assumptions (implicitly) define a hypothesis space $\mathcal{H}$, the space of hypotheses that could possibly be learned by the given algorithm.

Classification Algorithms and their Hypothesis-Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- These assumptions (implicitly) define a hypothesis space $\mathcal{H}$, the space of hypotheses that could possibly be learned by the given algorithm.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Hypothesis Space and Bias

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The hypothesis space is a manifestation of the so-called “bias”.
- Some algorithm, due to the restrictions of its hypothesis space, will prefer certain hypotheses, i.e., the algorithm has a bias.

- In general, we do not know if $f \in \mathcal{H}$, i.e., whether or not we can actually approximate f (well enough) with some (or any) specific classification algorithm.
- Note that a hypothesis space can be infinite and yet restricted.

Example: Some Concept of Enjoying Sport

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$f\colon \text{Sky} \times \text{AirTemp} \times \text{Humidity} \times \text{Wind} \times \text{Water} \times \text{Forecast} \rightarrow \{\text{Yes}, \text{No}\}$

- Is there a general concept when to enjoy sport?
- To learn a concept from the example data, we need some assumption on this concept.
- Here we can assume, for example, that the concept is a conjunction of some attribute values.
- We thus have a finite hypothesis space, containing all possible conjunctions of concrete attribute values.

Coverage of Data Instances by Hypotheses

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For this toy example, we define:

Example An example or instance x is a possible day, described as $x \in$ Sky $\times$ AirTemp $\times$ Humidity $\times$ Wind $\times$ Water $\times$ Forecast $= \mathcal{D}$

Positive/negative example An example x is positive, if $f(x) =$ Yes, negative otherwise.
(Note that we know f on the training data only.)

Hypothesis A hypothesis defines a subset of $\mathcal{D}$.
We write a hypothesis as vector containing

- specific values for attributes
- wildcards ('?') indicating that for some attribute any attribute value is acceptable
- $\emptyset$ indicating that no attribute value is acceptable.

Coverage of Data Instances by Hypotheses

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- For example the hypothesis $\langle \text{Sunny}, ?, ?, ?, \text{Cool}, ? \rangle$ defines all examples where Sky=Sunny and Water=Cool.
- If we are interested in hypotheses about positive examples, we can interpret $\langle \text{Sunny}, ?, ?, ?, \text{Cool}, ? \rangle$ as a rule:

```
if Sky=Sunny and Water=Cool
then EnjoySport=Yes
```

- An example x *satisfies* a hypothesis h, if and only if $f(x) = \text{Yes}$.

The Assumptions of a Learning Algorithm Define the Hypothesis Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$x_1 = \langle$Sunny, Warm, High, Strong, Cool, Same$\rangle$

$x_2 = \langle$Sunny, Warm, High, Light, Warm, Same$\rangle$

$h_1 = \langle$Sunny, ?, ?, Strong, ?, ?$\rangle$

$h_2 = \langle$Sunny, ?, ?, ?, ?, ?$\rangle$

$h_3 = \langle$Sunny, ?, ?, ?, Cool, ?$\rangle$

Definition 5.1

For any two hypotheses, h_j and h_k, over X, h_j is *more general than or equal to* h_k if and only if:

$$\forall x \in X : h_k(x) = \text{Yes} \Rightarrow h_j(x) = \text{Yes}$$

Basic Algorithm

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 5.1 (Find-S [Mitchell, 1997])

1. *Initialize h to the most specific hypothesis in $\mathcal{H}$*
2. *For each positive training instance x*
   For each attribute constraint a_i in h
   If the constraint a_i is satisfied by x
   Then do nothing
   Else replace a_i in h by the next more general constraint that is satisfied by x
3. *output hypothesis h*

Discussion:

- Finds the most specific hypothesis consistent with the positive examples in the training data – is it the only consistent hypothesis?
- Why should we prefer more specific hypotheses over more general ones?

Possible Hypotheses Under the Assumption of a Conjunctive Concept

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

The following conjunctive hypotheses describe the concept "Yes" correctly for the training data:

$\langle \text{Sunny}, ?, ?, ?, ?, ? \rangle$

$\langle ?, \text{Warm}, ?, ?, ?, ? \rangle$

$\langle \text{Sunny}, \text{Warm}, ?, ?, ?, ? \rangle$

$\langle \text{Sunny}, ?, ?, \text{Strong}, ?, ? \rangle$

$\langle ?, \text{Warm}, ?, \text{Strong}, ?, ? \rangle$

$\langle \text{Sunny}, \text{Warm}, ?, \text{Strong}, ?, ? \rangle$

Reduce the Bias?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For some other training data, our model assumptions seem too strict:

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Cool	Change	Yes
Cloudy	Warm	Normal	Strong	Cool	Change	Yes
Rainy	Warm	Normal	Strong	Cool	Change	No

- No consistent hypothesis is possible under our assumptions:
- the most specific hypothesis for positive examples is $\langle ?, \text{Warm}, \text{Normal}, \text{Strong}, \text{Cool}, \text{Change} \rangle$
- this hypothesis covers also the negative example

Bias-free Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Should we allow for more complex/generous assumptions to reduce the bias (i.e., to get rid of restrictions of the hypothesis space)?
- Allow disjunctions? Negations?
- A disjunctive hypothesis
  "if Sky=Sunny *or* Sky=Cloudy, then Yes"
  can list all positive examples.
- We could actually cover *any* concept with an arbitrarily complex hypothesis.
- Is this what we want?

Bias-free Learning is Futile

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Being able to cover *any* concept means, we don't rely on any assumption regarding possible concepts.
- The most complex hypothesis would describe the training data perfectly well: we can learn by heart all the examples.
- Learning by heart means we can perfectly predict all classes correctly for the *training examples* but we don't have any concept to *generalize to unseen data*.
- In general, a bias is necessary to avoid learning by heart.

Recommended Reading:

Mitchell [1997], Ch. 2

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The Problem of Evaluation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A database $\mathcal{D}$ is representing a domain by the sample of available data, there is a function $f : \mathcal{D} \rightarrow C$, mapping each object to a class $c_i \in C$.
- $O \subseteq \mathcal{D}$ is the set of objects where we know about the class (i.e., we know $f(o)$ for all $o \in O$, but not for any $o \in \mathcal{D} \setminus O$).
- Let h be a classifier (model, hypothesis), trained on a training set $TR \subseteq O$.

Problem:

- We want a good quality (performance, approximation of the target function f) of h over $\mathcal{D}$, yet we cannot know anything about the quality of h over $\mathcal{D} \setminus O$.

The Problem of Evaluation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- h has been optimized for TR.
- If we test h on TR, we will typically get an optimistic estimate of the performance over $\mathcal{D}$.
- The phenomenon that h performs better on TR than on $\mathcal{D}$ overall is called *overfitting*:
  - Often, the amount of training data is not sufficient to generalize reliably ("small sample size bias").
  - If the data sample is too small to truly represent the domain, there is no sufficient ground to reject overspecialized hypotheses.
  - In general, the weaker the bias of a learning algorithm, the more susceptible is it to overfitting.

The Separation of Training and Testing Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

To get a more realistic estimate of the performance of some classifier over $\mathcal{D}$, we separate training data $TR \subset O$ and test data $TE \subset O$, $TR \cap TE = \emptyset$.

- TR is used to train the classifier (fit the model, select the hypothesis).
- TE is used to evaluate the classifier.
- The purpose is to estimate the performance (success-/error-rate) of the classifier. Therefore:
  - TR and TE need to be independent ($TR \cap TE = \emptyset$).
  - Both, TR and TE should represent the classification problem as good as possible.
  - For some learning problems (benchmark datasets), a separation into TR and TE is available.
- Problem: If it is not clear whether O is already too small to allow for a good generalization, we do not really want TR to be even smaller.

m-fold Cross-Validation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Motivation: We want to use as much data as possible for training, and as much as possible for testing.

Algorithm 5.2 (m-fold cross-validation)

1. *Separate the set O in m equal-size, mutually disjoint subsets.*
2. *Get m different pairs of TR and TE by using each of the m subsets as TE once and the remaining $m-1$ subsets for training.*
3. *On these m pairs of TR and TE, train and test m independent classifiers.*
4. *Average the m observed performances.*
5. *Repeat 1-4 several times.*

m-fold Cross-Validation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For one iteration with $m = 3$, we have T_1 with class information a, T_2 with class information b, and T_3 with class information c:

Stratification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Stratification aims at representing the class proportions in each fold.

- minimum requirement: each class should be present in the training set.
- stratified cross-validation: the distribution of classes within each training and test set should reflect the distribution of the classes in $\mathcal{O}$

Standard approach (rule of thumb):
10-fold, stratified cross-validation, repeated 10 times

Note that:

The evaluation procedure has the purpose of estimating the performance on $\mathcal{D} \setminus \mathcal{O}$. In order to get the best possible classifier, we would use all available labeled data ($\mathcal{O}$) for training.

The Bootstrap

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

In the bootstrap procedure, a training set is created from O by drawing with replacement:

- We take $|O|$ objects from O, where the same object could be drawn several times.
- A sample TR contains on average 63% of the objects in O. Some are present several times in TR, some ($\approx 37\%$) are not present at all:
  - For $|O| = n$, an individual object in O has a chance of being drawn of $\frac{1}{n}$ each turn, that is, it is *not* drawn with probability $1 - \frac{1}{n}$.
  - After n draws, a specific object has not been drawn with probability $\left(1 - \frac{1}{n}\right)^n$
  - For large n: $\left(1 - \frac{1}{n}\right)^n \approx e^{-1} \approx 0.368$, hence this procedure is also called “the 0.632 bootstrap”

Leave-one-out or Jackknife Test

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 5.3 (Leave-one-out)

- $\forall o_j \in O$: *take* o_j *as test object for a classifier trained on* $O \setminus \{o_j\}$.
- *average the performance estimate over all test objects*

Discussion:

- For $|O| = n$, this is an n-fold cross-validation.
- Pro: no random effect
- Con: stratification is not possible
- In general, the Jackknife test leads to a relatively pessimistic performance estimate.

Confusion Matrix

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The confusion matrix represents the number of correctly and incorrectly predicted classes per actual and per predicted class:

predicted class

actual class

	class 1	class 2	class 3	class 4	class 5
class 1	35	1	1	1	4
class 2	0	31	1	1	5
class 3	3	1	50	1	2
class 4	1	0	1	10	2
class 5	3	1	9	15	13

correctly predicted objects

Quality Measures for Classifiers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given a classifier h, a training set $TR \subseteq O$, and a test set $TE \subseteq O$. $f(o)$ is the actual class of o, $h(o)$ is the class predicted by the classifier h. Then we have:

accuracy of h on TE:

$$\mathrm{acc}_{TE}(h) = \frac{|\{o \in TE | h(o) = f(o)\}|}{|TE|}$$

true classification error:

$$\mathrm{err}_{TE}(h) = \frac{|\{o \in TE | h(o) \neq f(o)\}|}{|TE|}$$

apparent classification error:

$$\mathrm{err}_{TR}(h) = \frac{|\{o \in TR | h(o) \neq f(o)\}|}{|TR|}$$

Focus on Individual Classes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If we focus on a single class (the "positive" class) vs. all other classes (the "negative" class), the confusion matrix can be read as follows:

	predicted positive	predicted negative
given positive	TP (true positive)	FN (false negative)
given negative	FP (false positive)	TN (true negative)

This notation is also often used in two-class problems, where we have a particular interest to detect cases of the "positive" class, e.g., medical tests on specific diseases.

Quality Measures for Individual Classes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

recall: proportion of test objects of some class c_i that have been predicted correctly

$$f_i = \{o \in TE | f(o) = c_i\} :$$

$$\text{recall}_{TE}(h, i) = \frac{|\{o \in f_i | h(o) = f(o)\}|}{|f_i|}$$

precision: proportion of test objects predicted as class c_i that actually belong to class c_i

$$h_i = \{o \in TE | h(o) = c_i\} :$$

$$\text{precision}_{TE}(h, i) = \frac{|\{o \in h_i | h(o) = f(o)\}|}{|h_i|}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Instance-based Learning

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- a simple classifier: assign to a new object the class of the nearest neighbor in the training data

- we can visualize class regions by Voronoi cells

k-Nearest Neighbor Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- potential problem: nearest neighbor might be an outlier, somehow unusual, misleading
- take more than one neighbor into consideration: k nearest neighbor classifier

decision set: the set of (k) nearest neighbors considered for the classification decision

decision rule: how to decide the class, given the potentially different classes of the k nearest neighbors
- take the majority vote
- potentially weighted votes

Let $x_1, \ldots, x_k$ be the k nearest neighbors of instance x_q:

$$h(x_q) = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \delta(c, f(x_i)) \text{ where } \delta(a,b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise} \end{cases}$$

"Training" and Prediction for kNN Classifier

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

small training set: simple decision boundaries

"Training" and Prediction for kNN Classifier

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

small training set: simple decision boundaries

"Training" and Prediction for kNN Classifier

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

large training set: potentially complex decision boundaries

"Training" and Prediction for kNN Classifier

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

large training set: potentially complex decision boundaries

Complex Class Boundaries with Larger k

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

dataset

Complex Class Boundaries with Larger k

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$k = 1$

Complex Class Boundaries with Larger k

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$k = 5$

Decision Set

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- k too small: classifier is sensitive to outliers
- k too large: potentially takes objects belonging to other classes intot he decision set
- medium k: best quality
- rule of thumb: $1 \ll k \leq 10$, but consider, e.g., size of classes in training set

Decision Rule

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- standard: choose majority class in decision set
- advanced: put weights on the class votes
  - by distance, typically squared inverted distance:

$$\text{weight}(\text{dist}) = \frac{1}{\text{dist}^2}$$

  - by class proportions:
    If a class is small, it has a smaller chance of being the majority in some decision set.

Example: 2 classes, 95% A, 5% B, the labels of some decision set (e.g., labels of the 7 nearest neighbors of x) are $\{A, A, A, A, B, B, B\}$

- standard decision rule: $h(x) = A$
- votes weighted based on class size: $h(x) = B$

Instance-based Learning

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Instance-based learning does not provide an explicit description of the target function.
- Training examples are simply stored.
- Generalization beyond the training examples is postponed until a new instance must be classified ("lazy learner").
- High flexibility (low bias) because the target function is actually estimated locally and differently for each new instance instead of once for the entire instance space.

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

pros

- easy to apply: requires only training data and distance function
- often good classification accuracy
- incremental: easy adaptation to new training data
- no training required ("lazy learner")

cons

- inefficient prediction: each decision requires k nearest neighbor query
- does not deliver explicit knowledge about classes
- difficult to choose k, esp. if classes are of very different size

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
- Introduction
- Bias-free Learning?
- Evaluation of Classifiers
- k-Nearest Neighbor Classification
- Summary

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Introduction
Bias-free Learning?
Evaluation of Classifiers
k-Nearest Neighbor Classification
Summary
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *What is Classification?*
- *hypothesis-space and bias*
- *Why is a bias unavoidable for learning and generalization?*
- *evaluation procedures (cross-validation, bootstrap, leave-one-out)*
- *quality measures for classifiers:*
  - *confusion matrix*
  - *accuracy & error (apparent vs. true)*
  - *precision & recall*
- *a simple classifier: k-nearest neighbors*
  - *properties, challenges, variants*
  - *lazy learning*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Can we Formalize "Confidence"?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

We encountered a measure called "confidence" that should tell us how reliable a discovered association rule is.
We interpreted:

- The higher the confidence for some rule '$X \Rightarrow Y$', the more likely Y is present in transactions that contain X.
- The confidence is an estimate of the conditional probability of Y given X.

Can we formalize this interpretation?

Recommended Reading:

Mitzenmacher and Upfal [2017], Chapter 1.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Sample Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The sample space Ω is the set of all (disjoint) possible outcomes of some random process.

Examples:

- *If we role a dice, we have* $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- *If we flip a coin, we have* $\Omega = \{H, T\}$.

Events

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A subset $E \subseteq \Omega$ of individual outcomes of a random process can define an "event".

Examples:

- *We role a die. Every element of* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *is a simple or elementary event.*
- *We could be interested in the event "The die shows an even number"* $= \{2, 4, 6\} \subseteq \Omega$.
- *We flip a coin. We could have the elementary event "head"* $\subseteq \Omega$.

A family of sets $\mathcal{F}$ represents the allowable events. Each set in $\mathcal{F}$ is a subset of Ω, i.e., $\mathcal{F} \subseteq \wp(\Omega)$.

Probability Function

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 6.1 (Probability Function)

A probability function is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies the following conditions:

1. $\forall E : 0 \leq \Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$; and
3. for any finite or countably infinite sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$:

$$\Pr\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} \Pr(E_i).$$

Probability Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 6.2 (Probability Space)

A probability space is given by three components:

1. a sample space Ω;
2. the allowable events $\mathcal{F} \subseteq \wp(\Omega)$; and
3. a probability function $\mathrm{Pr} : \mathcal{F} \to \mathbb{R}$.

Event Combinations

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Because events are sets, we can use standard set theory notation to express combinations of events.

- $E_1 \cap E_2$ denotes the occurrence of both, E_1 and E_2 (i.e., their co-occurrence).
- $E_1 \cup E_2$ denotes the occurrence of either E_1 or E_2 (or both).
- $E_1 \setminus E_2$ denotes the occurrence of event E_1 without E_2 occurring as well.
- $\overline{E} = \Omega \setminus E$ denotes the complementary event of E.

Event Combinations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Examples:

Suppose we roll two dice. Given events E_1 and E_2:

E_1 *the first die is a 1*

E_2 *the second die is a 1*

- $E_1 \cap E_2$: *both dice are 1*
- $E_1 \cup E_2$: *at least one of the dice lands on 1.*
- $E_1 \setminus E_2$: *the first die is a 1 and the second die is not.*

Event Combinations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Examples:

Let E be the event that by rolling a die we obtain an even number.

- *Then $\overline{E}$ is the event that we obtain an odd number.*
- *What are the events $\overline{E_1}$, $\overline{E_1 \cup E_2}$, $\overline{E_1 \cap E_2}$?*

Example: One Die

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling a die:

$$\Omega_{\text{die}_1} = \{1, 2, 3, 4, 5, 6\}$$

Assuming a fair die, all sides have equal probability, thus:

$$\Pr(\{1\}) = \Pr(\{2\}) = \ldots = \Pr(\{6\}) = \frac{1}{6}$$

The probability of the event "odd number" is

$$\Pr(\{1, 3, 5\}) = \Pr(\{1\}) + \Pr(\{3\}) + \Pr(\{5\}) = \frac{1}{2}$$

Example: Two Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event "sum $= 2$":

$$\Pr(\{(1,1)\}) = \frac{1}{36}$$

Example: Two Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event "sum $= 3$":

$$\Pr(\{(1,2),(2,1)\}) = \Pr(\{(1,2)\}) + \Pr(\{(2,1)\}) = \frac{2}{36} = \frac{1}{18}$$

Example: Two Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event E_1 ="sum bounded by 6":

$$E_1 = \{(1,1),(1,2),(1,3),(1,4),(1,5),(2,1),(2,2),$$
$$(2,3),(2,4),(3,1),(3,2),(3,3),(4,1),(4,2),(5,1)\}$$

$$\Pr(E_1) = \frac{15}{36}$$

Example: Two Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

$E_2 =$ *"both dice have odd numbers":*

$$E_2 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$$

$$\Pr(E_2) = \frac{1}{4}$$

Example: Two Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

$E_3 =$ *"sum bounded by 6 and both dice have odd numbers":*

$$\begin{aligned}\Pr(E_3) &= \Pr(E_1 \cap E_2) \\ &= \Pr(\{(1,1),(1,3),(1,5),(3,1),(3,3),(5,1)\}) \\ &= \frac{1}{6}\end{aligned}$$

Combined Probability

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 6.3 (Combined Probability)

For any two events E_1 and E_2:

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

Proof.

$$\begin{aligned}
\Pr(E_1) &= \Pr(E_1 \setminus (E_1 \cap E_2)) + \Pr(E_1 \cap E_2) \\
\Pr(E_2) &= \Pr(E_2 \setminus (E_1 \cap E_2)) + \Pr(E_1 \cap E_2) \\
\Pr(E_1 \cup E_2) &= \Pr(E_1 \setminus (E_1 \cap E_2)) \\
&\quad + \Pr(E_2 \setminus (E_1 \cap E_2)) \\
&\quad + \Pr(E_1 \cap E_2)
\end{aligned}$$

□

Union Bound

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 6.4 (Union Bound)

For any finite or countably infinite sequence of events $E_1, E_2, E_3, \ldots$*:*

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) \leq \sum_{i\geq 1} \Pr(E_i).$$

Difference from condition 3 in Definition 6.1?

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
  - Quality Measures for Association Rules Revisited
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Independent Events

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 6.5 (Independent Events)

Two events E and F are *independent* if and only if

$$\Pr(E \cap F) = \Pr(E) \cdot \Pr(F).$$

More generally, events $E_1, E_2, \ldots, E_k$ are mutually independent if and only if

$$\forall I \subseteq [1,k] : \Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

Independent Events: Intuition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- If events A and B are *independent* then knowledge about event A does not change the probability of B.
- If A and B are *not independent*, then we can quantify the conditional probability of A subject to our kowledge of event B.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 4": $\frac{4}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 4":

$$\Pr(E_1 \cap E_2) = \frac{2}{6} = \frac{12}{36} = \frac{3}{6} \cdot \frac{4}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are independent.*

Independent Events: Intuition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- If events A and B are *independent* then knowledge about event A does not change the probability of B.
- If A and B are *not independent*, then we can quantify the conditional probability of A subject to our kowledge of event B.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 3": $\frac{3}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 3":

$$\Pr(E_1 \cap E_2) = \frac{1}{6} = \frac{6}{36} \neq \frac{9}{36} = \frac{3}{6} \cdot \frac{3}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are not independent.*

Conditional Probability

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 6.6 (Conditional Probability)

The *conditional probability* that event E occurs given that event F occurs is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

The conditional probability is well-defined only if $\Pr(F) > 0$.

Note that:

If E and F are independent and $\Pr(F) \neq 0$, we have:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E)$$

Conditional Probability: Intuition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We look for the probability of $E \cap F$ within the sets of events defined by F.
- Because F restricts the sample space, we normalize the probabilities by dividing by $\Pr(F)$.
- If E and F are independent, information about F should not affect the probability of E.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 4": $\frac{4}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 4":

$$\Pr(E_1 \cap E_2) = \frac{2}{6} = \frac{12}{36} = \frac{3}{6} \cdot \frac{4}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are independent.*

Conditional Probability: Intuition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We look for the probability of $E \cap F$ within the sets of events defined by F.
- Because F restricts the sample space, we normalize the probabilities by dividing by $\Pr(F)$.
- If E and F are independent, information about F should not affect the probability of E.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 3": $\frac{3}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 3":

$$\Pr(E_1 \cap E_2) = \frac{1}{6} = \frac{6}{36} \neq \frac{9}{36} = \frac{3}{6} \cdot \frac{3}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are not independent.*

The Condition Defines a Probability Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\Pr(X|E)$ defines a proper probability function on the sample space E (cf. Definitions 6.1 and 6.2):

$$\Pr(\emptyset|E) = \frac{\Pr(\emptyset \cap E)}{\Pr(E)} = \frac{\Pr(\emptyset)}{\Pr(E)} = 0$$

$$\Pr(E|E) = \frac{\Pr(E \cap E)}{\Pr(E)} = \frac{\Pr(E)}{\Pr(E)} = 1$$

For any two disjoint events A and B:

$$\begin{aligned}\Pr(A \cup B|E) &= \frac{\Pr((A \cup B) \cap E)}{\Pr(E)} \\ &= \frac{\Pr(A \cap E) + \Pr(B \cap E)}{\Pr(E)} \\ &= \Pr(A|E) + \Pr(B|E)\end{aligned}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
  - Quality Measures for Association Rules Revisited
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Quality Measures for Association Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Support: $s(X \Rightarrow Y) = s(X \cup Y)$

or in relative terms: frequency $f(X \cup Y) = \frac{s(X \cup Y)}{|\mathcal{D}|}$

Confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Lift: $\mathit{Lift}(X \Rightarrow Y) = \frac{\mathrm{conf}(X \Rightarrow Y)}{f(Y)}$

Jaccard: $\mathit{Jaccard}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)+s(Y)-s(X \cup Y)}$

conviction: $\mathit{conviction}(X \Rightarrow Y) = \frac{1-f(Y)}{1-\mathrm{conf}(X \Rightarrow Y)}$

Probabilistic Interpretation: Support (Frequency)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The frequency of an itemset in the database can be seen as an empirical estimate of its probability, given the sample represented by the database:

$$\Pr(X) = \frac{s(X)}{|\mathcal{D}|}$$

Note that:

$$\Pr(X \cap Y) = \frac{s(X \cup Y)}{|\mathcal{D}|}$$

Although X and Y are sets in both cases,

- *probabilistically, $\cap$ denotes the co-occurrence of events,*
- *while for itemsets, $\cup$ denotes that both itemsets need to be present.*

Probabilistic Interpretation: Confidence

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The confidence is the conditional probability that a transaction contains the consequent Y given that it contains the antecedent X:

$$\begin{aligned}\mathrm{conf}(X \Rightarrow Y) &= \frac{s(X \cup Y)}{s(X)} \\ &= \frac{\Pr(X \cap Y)}{\Pr(X)} \\ &= \Pr(Y|X)\end{aligned}$$

- The confidence of a rule $X \Rightarrow Y$ is not a useful measure unless we compare it with the frequency of Y, i.e., the prior (unconditional) probability.
- If we have $\Pr(Y|X) < \Pr(Y)$ this means that in the presence of X, Y becomes less likely as it is unconditionally!
  (Not the rule, but this fact could be interesting, though!)

Probabilistic Interpretation: Lift

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

We can see Lift as normalization of the confidence by the prior probability of the consequent:

$$\begin{aligned} \mathit{Lift}(X \Rightarrow Y) &= \frac{\mathrm{conf}(X \Rightarrow Y)}{f(Y)} \\ &= \frac{\Pr(X \cap Y)}{\Pr(X)\Pr(Y)} \end{aligned}$$

- ratio of the observed joined probability of X and Y to the joint probability expected for statistically independent events (Definition 6.5).
- Lift is a (symmetric!) measure for the surprise of a rule.
- Values around 1: boring.
- Much smaller/larger values: interesting!

Probabilistic Interpretation: Jaccard

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The Jaccard coefficient in general is a measure for the similarity between two sets:

$$\begin{aligned} \mathit{Jaccard}(X \Rightarrow Y) &= \frac{s(X \cup Y)}{s(X) + s(Y) - s(X \cup Y)} \\ &= \frac{\Pr(X \cap Y)}{\Pr(X) + \Pr(Y) - \Pr(X \cap Y)} \\ &= \frac{\Pr(X \cap Y)}{\Pr(X \cup Y)} \qquad \text{(Lemma 6.3)} \end{aligned}$$

- A symmetric measure of how often both, X and Y, occur simultaneously, relative to the occurrence of both or either overall.
- Similarity of the itemsets X and Y based on their individual occurrences and their co-occurrences.

Probabilistic Interpretation: Conviction

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Conviction of a rule measures the expected error: how often does X occur in a transaction where Y does not?
(How often does the rule fail?)

$$\begin{aligned}
\textit{conviction}(X \Rightarrow Y) &= \frac{1 - f(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)} \\
&= \frac{\Pr\left(\overline{Y}\right)}{1 - \frac{\Pr(X \cap Y)}{\Pr(X)}} = \frac{\Pr(X)\Pr\left(\overline{Y}\right)}{\Pr(X) - \Pr(X \cap Y)} \\
&= \frac{\Pr(X)\Pr\left(\overline{Y}\right)}{\Pr\left(X \cap \overline{Y}\right)} = \frac{1}{\textit{Lift}\left(X \Rightarrow \overline{Y}\right)}
\end{aligned}$$

- compares the observed joint probability of X and $\overline{Y}$ with their joint probability expected for independence
- asymmetric measure

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

On the probabilistic interpretation of (even more) quality measures for association rules: Zaki and Meira Jr. [2014], Chapter 12.1.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The Law of Total Probability

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 6.7 (The Law of Total Probability)

Let $E_1, E_2, \ldots, E_n$ be mutually disjoint events in the sample space Ω, and let $\bigcup_{i=1}^{n} E_i = \Omega$. Then

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B|E_i)\Pr(E_i).$$

The Law of Total Probability

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Proof.

Since the events $E_i (i = 1, \ldots, n)$ are disjoint and cover the entire sample space Ω, it follows that

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i).$$

Further, by Definition 6.6,

$$\sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B|E_i) \Pr(E_i).$$

□

Bayes' Rule

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 6.8 (Bayes' Rule)

Let $E_1, \ldots, E_n$ be mutually disjoint events, and let $\bigcup_{i=1}^{n} E_i = \Omega$. Then for any other event B, $\Pr(B) > 0$, $j = 1, \ldots, n$:

$$\Pr(E_j|B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} \tag{6.1}$$

$$= \frac{\Pr(B|E_j)\Pr(E_j)}{\sum_{i=1}^{n} \Pr(B|E_i)\Pr(E_i)} \tag{6.2}$$

Proof.

From Eq. 1 to Eq. 2, we use Definition 6.6 in the numerator, and Theorem 6.7 in the denominator. □

Bayes' Rule (Simple Form)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

In its simple form, we have only two events, A and B, $\Pr(A) \neq 0$, $\Pr(B) \neq 0$:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)}$$

$$\Rightarrow \Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

$$\Rightarrow \Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

We do not require exhaustiveness of A or B here (i.e., $A \subseteq \Omega$, $B \subseteq \Omega$), since we do not apply Theorem 6.7, only Definition 6.6.

Bayes' Rule: Example 1

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We are given three coins, two of the coins are fair and the third coin is biased, showing head with probability $\frac{2}{3}$. We need to identify the biased coin.
- We flip each of the coins. The first and second coins come up with head, the third coin comes up with tail.
- What is the probability that the first coin is the biased one?
- Let E_i be the event that the i-th coin is the biased one, and let B be the event that the three coin flips came up head, head, tail.
- Prior probability: $\Pr(E_i) = \frac{1}{3}$ for $i = 1, 2, 3$.

Bayes' Rule: Example 1

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\Pr(B|E_1) = \Pr(B|E_2) = \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{6}$$

and

$$\Pr(B|E_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{12}.$$

Thus, according to Bayes' rule:

$$\Pr(E_1|B) = \frac{\Pr(B|E_1)\Pr(E_1)}{\sum_{i=1}^{3} \Pr(B|E_i)\Pr(E_i)} = \frac{2}{5}.$$

The experiment increases the probability that the first coin is the biased one from $\frac{1}{3}$ to $\frac{2}{5}$.

Bayes' Rule: Example 2

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A doctor sees a patient with fever and rash.
- 80% of patient with flu, 45% of allergy patients, and 90% of infection patients have these symptoms.
- The doctor knows that 50% of the patients she sees have flu, 40% have allergy, and 10% have an infection.
- Should the doctor treat the patient for infection?

Bayesian Reasoning: The General Pattern

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- There are alternative models to explain a fact.
- Each model defines a probability for the observed data.
- Which model is the best (i.e., the most likely) explanation?
- $E_1, E_2, \ldots, E_n$ are the alternative models.
- B is the observed data.
- For each model we know $\Pr(B|E_j)$ (i.e., how well the model explains the facts).

$$\Pr(E_j|B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} = \frac{\Pr(B|E_j)\Pr(E_j)}{\sum_{i=1}^{n} \Pr(B|E_i)\Pr(E_i)}$$

- Difficulty: How do we know $\Pr(E_j)$?

Bayesian Approach

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Start with a *prior* model, giving some initial value to the model parameters.
- This model is then modified by incorporating new observations, to obtain a *posterior* model that captures the new information.

Example:

- A test shows that a patient has an infection.
- The test has 10% error rate.
- What is the probability that the patient has an infection?

Bayesian Approach: Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $A =$ "test is positive (i.e., the test says that the patient has an infection)"
- $B =$ "the patient actually has an infection"

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr\left(A|\overline{B}\right)\Pr\left(\overline{B}\right)}$$

- What is $\Pr(B)$?

Without any prior knowledge we set $\Pr(B) = \Pr\left(\overline{B}\right) = \frac{1}{2}$:

$$\Pr(B|A) = \frac{\frac{9}{10} \cdot \frac{1}{2}}{\frac{9}{10} \cdot \frac{1}{2} + \frac{1}{10} \cdot \frac{1}{2}} = \frac{9}{10}$$

The estimate is dominated by the reliability of the test.

Bayesian Approach: Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $A =$ "test is positive (i.e., the test says that the patient has an infection)"
- $B =$ "the patient actually has an infection"

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr\left(A|\overline{B}\right)\Pr\left(\overline{B}\right)}$$

- What is $\Pr(B)$?

Assume that we know a priori that the probability of the patient being infected is 80%. We set $\Pr(B) = \frac{4}{5}$:

$$\Pr(B|A) = \frac{\frac{9}{10} \cdot \frac{4}{5}}{\frac{9}{10} \cdot \frac{4}{5} + \frac{1}{10} \cdot \frac{1}{5}} = \frac{36}{37} \approx 0.97$$

The *posterior* probability is sensitive to the choice of *prior* probabilities.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Mitchell [1997], Chapter 6.*
- *Zaki and Meira Jr. [2014], Chapter 18.*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Probabilistic Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A classifier predicts for some object x_q which class c_i it belongs to.
- Often, the prediction can be expressed as probability estimate
$$\Pr(c_i|x_q)$$
- The classifier would then decide for the most likely class:
$$h(x_q) = \arg\max_{c_i \in C} \Pr(c_i|x_q)$$
- Often, this estimate is based on an estimate of how likely the object would be, if it would belong to this or to that class:
$$\Pr(x_q|c_i)$$
How well can the object be explained if it belongs to a given class?

Proportions of Classes in the Decision Set

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Let $x_1, \ldots, x_k = kNN(x_q)$ be the k nearest neighbors of instance x_q, i.e., the decision set for the instance x_q.
- For a given instance x_q and classes $C = \{c_1, \ldots, c_m\}$:

$$E_j = \text{“}f(x_q) = c_j\text{”}$$

$$\Omega = \bigcup_{j=1}^{m} E_j$$

- The relative frequency of a class c_j in the decision set $kNN(x_q)$ is an empirical estimate of the probability of the event “$f(x_q) = c_j$”:

$$\Pr(E_j|x_q) = \frac{|\{x_i | x_i \in kNN(x_q) \wedge f(x_i) = c_j\}|}{k}$$

$$= \frac{\sum_{i=1}^{k} \delta(c_j, f(x_i))}{k} \text{ with } \delta(a,b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise} \end{cases}$$

Proportions of Classes in the Decision Set

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We can therefore interpret the composition of the decision set as "class probability vector".
- For classes $C = \{c_1, \ldots, c_m\}$, the decision set for x_q yields a vector

$$\langle p_1, \ldots, p_m \rangle$$

where

$$\begin{aligned} p_j &= \Pr\left(\{f(x_q) = c_j\} | x_q\right) \\ &= \frac{|\{x_i | x_i \in kNN(x_q) \wedge f(x_i) = c_j\}|}{k} \\ &= \frac{\sum_{i=1}^{k} \delta(c_j, f(x_i))}{k} \end{aligned}$$

- How will the quality of the probability estimate depend on k?

Decision Rule — Weight by Distance

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

As discussed (Slides 231, 235), we can introduce a weight to the components in the decision rule:

$$h(x_q) = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \delta(c, f(x_i))$$

- e.g.: $w_i = \frac{1}{\mathrm{dist}(x_i, x_q)^2}$

$$\Pr_w(\{f(x_q) = c_j\} | x_q) = \frac{\sum_{i=1}^{k} w_i \delta(c_j, f(x_i))}{\sum_{i=1}^{k} w_i}$$

- How do these weights change the dependency on k?

kNN Classifier as Application of Bayes' Rule

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Let $k_i = |\{x | x \in kNN(x) \wedge f(x) = c_i\}|$.
- Let $n_i = |\{x | x \in O \wedge f(x) = c_i\}|$, i.e., $\Pr(c_i) = \frac{n_i}{|O|}$
- Let $V_k(x)$ be the volume of the $kNN(x)$.
- $\Pr(x|c_i) = \frac{\frac{k_i}{n_i}}{V_k(x)} = \frac{k_i}{n_i \cdot V_k(x)}$
- $\Pr(x|c_i) \cdot \Pr(c_i) = \frac{k_i}{n_i \cdot V_k(x)} \cdot \frac{n_i}{|O|} = \frac{k_i}{|O| \cdot V_k(x)}$

$$\Pr(c_i|x) = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\sum_{j=1}^{m} \Pr(x|c_j) \cdot \Pr(c_j)}$$

$$= \frac{\frac{k_i}{|O| \cdot V_k(x)}}{\sum_{j=1}^{m} \frac{k_j}{|O| \cdot V_k(x)}} = \frac{k_i}{k}$$

$$h(x) = \arg\max_{c_i \in C} (\Pr(c_i|x)) = \arg\max_{c_i \in C} \left(\frac{k_i}{k}\right) = \arg\max_{c_i \in C}(k_i)$$

Decision Rule — Weight by Prior Class Probability

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- To account for very imbalanced proportions of class sizes, we can adjust the prior probability.
- Because we *want* each class *a priori* to be equally likely, we set $\Pr(c_i) = \frac{1}{m}$.
- $\Pr(x|c_i) \cdot \Pr(c_i) = \frac{k_i}{n_i \cdot V_k(x)} \cdot \frac{1}{m} = \frac{k_i}{n_i \cdot m \cdot V_k(x)}$

$$\Pr(c_i|x) = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\sum_{j=1}^{m} \Pr(x|c_j) \cdot \Pr(c_j)}$$

$$= \frac{\frac{k_i}{n_i \cdot m \cdot V_k(x)}}{\sum_{j=1}^{m} \frac{k_j}{n_j \cdot m \cdot V_k(x)}} = \frac{\frac{k_i}{n_i}}{\sum_{j=1}^{m} \frac{k_j}{n_j}}$$

$$h(x) = \arg\max_{c_i \in C} (\Pr(c_i|x)) = \arg\max_{c_i \in C} \left(\frac{k_i}{n_i}\right)$$

- How does the decision change with the choice of k?

Decision Rule — Weight by Prior Class Probability (Example)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Given a training set O, $|O| = 100$, classes $C = \{A, B\}$, $n_A = 80$, $n_B = 20$
- We choose to set the prior probability $\Pr(A) = \Pr(B) = \frac{1}{2}$
- $k = 10$, classes of the $kNN(x)$ are: $\{A, A, A, A, A, A, B, B, B, B\}$, i.e., $k_A = 6$, $k_B = 4$

$$\Pr(A|x) = \frac{\Pr(x|A) \cdot \Pr(A)}{\Pr(x|A) \cdot \Pr(A) + \Pr(x|B) \cdot \Pr(B)}$$

$$= \frac{\frac{6}{80}}{\frac{6}{80} + \frac{4}{20}} = \frac{\frac{3}{40}}{\frac{3}{40} + \frac{8}{40}} = \frac{\frac{3}{40}}{\frac{11}{40}} = \frac{3}{11}$$

$$\Pr(B|x) = \frac{\frac{8}{40}}{\frac{11}{40}} = \frac{8}{11}$$

$$h(x) = \arg\max_{c_i \in C} (\Pr(c_i|x)) = \arg\max_{c_i \in C} \left(\frac{k_i}{n_i}\right)$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Probabilities and Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The aim of machine learning (or actually of science as such) could be put as "*find the best hypothesis to explain the observations*".
- If we approach learning probabilistically, "*best*" means "*most probable*", given the data $\mathcal{D}$ plus any initial knowledge about the prior probabilities of the various hypotheses in $\mathcal{H}$.
- The prior probability $\Pr(h)$ denotes the initial probability of hypothesis h before we observe the training data.
- The prior probability could reflect any background knowledge.
- The prior probability $\Pr(\mathcal{D})$ denotes the probability of the data (observations) without any knowledge on which hypothesis holds.

Prior and Posterior Probabilities, and Bayes' Theorem

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The conditional probability $\Pr(\mathcal{D}\,|h)$ denotes the probability of the observations (likelihood of the data), given some hypothesis h (i.e., assuming, h is correct).
- The probability $\Pr(h|\,\mathcal{D})$ is called the *posterior probability*, because it reflects our confidence that hypothesis h is correct *after* we have seen the training data $\mathcal{D}$.
- Given prior probabilities $\Pr(h)$, $\Pr(\mathcal{D})$, and conditional probability $\Pr(\mathcal{D}\,|h)$, the posterior probability $\Pr(h|\,\mathcal{D})$ can be computed by Bayes' theorem (Theorem 6.8):

$$\Pr(h|\,\mathcal{D}) = \frac{\Pr(\mathcal{D}\,|h) \cdot \Pr(h)}{\Pr(\mathcal{D})}$$

- Intuitive interpretation: $\Pr(h|\,\mathcal{D})$ increases with $\Pr(\mathcal{D}\,|h)$ and with $\Pr(h)$, it decreases as $\Pr(\mathcal{D})$ increases.

Maximum Likelihood and Maximum A Posteriori Hypothesis

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A classifier shall identify the most probable hypothesis $h \in \mathcal{H}$, given the observed data.
- We call such a maximally probable hypothesis a *maximum a posteriori* (MAP) hypothesis:

$$\begin{aligned} h_{\text{MAP}} &\equiv \arg\max_{h \in \mathcal{H}} \Pr(h|\,\mathcal{D}) \\ &= \arg\max_{h \in \mathcal{H}} \frac{\Pr(\mathcal{D}\,|h) \cdot \Pr(h)}{\Pr(\mathcal{D})} \\ &= \arg\max_{h \in \mathcal{H}} \Pr(\mathcal{D}\,|h) \cdot \Pr(h) \end{aligned}$$

- If we assume equal prior probabilities for all hypotheses (i.e., $\Pr(h_i) = \Pr(h_j) \forall h_i, h_j \in \mathcal{H}$), MAP is given by the maximum likelihood hypothesis (that maximizes the likelihood of the data):

$$h_{\text{ML}} \equiv \arg\max_{h \in \mathcal{H}} \Pr(\mathcal{D}\,|h)$$

Brute-Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If we have a finite hypothesis space $\mathcal{H}$, we can design a straightforward learning algorithm to find the maximum aposteriori hypothesis, based on Bayes' rule:

1. For each hypothesis $h \in \mathcal{H}$, calculate

$$\Pr(h|\,\mathcal{D}) = \frac{\Pr(\mathcal{D}\,|h)\Pr(h)}{\Pr(\mathcal{D})}$$

2. Return $h_{\text{MAP}} = \arg\max_{h\in\mathcal{H}} \Pr(h|\,\mathcal{D})$

Requires prior probabilities (estimates for $\Pr(h)$ and $\Pr(\mathcal{D}\,|h)$, $\Pr(\mathcal{D})$ will then be determined).

Assumptions for Brute-Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

To estimate the prior probabilities, we assume:

1. $\mathcal{D}$ is noise free (i.e., no incorrect labels, f is defined over $\mathcal{D}$).
2. The target concept f is contained in the hypothesis space $\mathcal{H}$.
3. We have no apriori reason to believe that any hypothesis is more probable than any other.

Assumptions for Brute-Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

To estimate the prior probabilities, we assume:

1. $\mathcal{D}$ is noise free (i.e., no incorrect labels, f is defined over $\mathcal{D}$).
2. The target concept f is contained in the hypothesis space $\mathcal{H}$.
3. We have no apriori reason to believe that any hypothesis is more probable than any other.

Assumption 3 $\Rightarrow$ same prior probability to every $h \in \mathcal{H}$.
Assumption 2 $\Rightarrow$ prior probabilities for all $h \in \mathcal{H}$ sum to 1.
Therefore:

$$\forall h \in \mathcal{H} : \Pr(h) = \frac{1}{|\mathcal{H}|}$$

Assumptions for Brute-Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

To estimate the prior probabilities, we assume:

1. $\mathcal{D}$ is noise free (i.e., no incorrect labels, f is defined over $\mathcal{D}$).
2. The target concept f is contained in the hypothesis space $\mathcal{H}$.
3. We have no apriori reason to believe that any hypothesis is more probable than any other.

Assumptions 1&2 $\Rightarrow$ If some h is correct (i.e., equals f on $\mathcal{D}$), it would explain the data perfectly.

$$\Pr(\mathcal{D}\,|h) = \begin{cases} 1 & \text{if } h(x_i) = f(x_i) \forall x_i \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Assumptions for Brute-Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

To estimate the prior probabilities, we assume:

1. $\mathcal{D}$ is noise free (i.e., no incorrect labels, f is defined over $\mathcal{D}$).
2. The target concept f is contained in the hypothesis space $\mathcal{H}$.
3. We have no apriori reason to believe that any hypothesis is more probable than any other.

If h is inconsistent with $\mathcal{D}$:

$$\Pr(h|\,\mathcal{D}) = \frac{0 \cdot \Pr(h)}{\Pr(\mathcal{D})} = 0$$

For $|\mathcal{H}_\mathcal{D}|$ consistent hypotheses in $\mathcal{H}$:

$$\Pr(h|\,\mathcal{D}) = \frac{1 \cdot \frac{1}{|\mathcal{H}|}}{\Pr(\mathcal{D})} = \frac{1 \cdot \frac{1}{|\mathcal{H}|}}{\frac{|\mathcal{H}_\mathcal{D}|}{|\mathcal{H}|}} = \frac{1}{|\mathcal{H}_\mathcal{D}|}$$

Prior Probability of $\mathcal{D}$ Under Assumptions for Brute Force MAP Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\Pr(\mathcal{D}) = \frac{|\mathcal{H}_\mathcal{D}|}{|\mathcal{H}|}$$

Proof.

By Theorem 6.7 and because the hypotheses are mutually exclusive (i.e., $\forall i \neq j : \Pr(h_i \wedge h_j) = 0$), we have:

$$\begin{aligned}\Pr(\mathcal{D}) &= \sum_{h_i \in \mathcal{H}} \Pr(\mathcal{D} \,|h_i) \Pr(h_i) \\ &= \sum_{h_i \in \mathcal{H}_\mathcal{D}} 1 \cdot \frac{1}{|\mathcal{H}|} + \sum_{h_i \notin \mathcal{H}_\mathcal{D}} 0 \cdot \frac{1}{|\mathcal{H}|} \\ &= \sum_{h_i \in \mathcal{H}_\mathcal{D}} 1 \cdot \frac{1}{|\mathcal{H}|} = \frac{|\mathcal{H}_\mathcal{D}|}{|\mathcal{H}|}\end{aligned}$$

□

MAP Hypotheses and Consistent Learners

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Result of our analysis: given our choices for $\Pr(h)$ and $\Pr(\mathcal{D}\,|h)$, every consistent hypothesis has posterior probability $\frac{1}{|\mathcal{H}_\mathcal{D}|}$, every inconsistent hypothesis has posterior probability 0.
- Therefore, every consistent hypothesis is an MAP hypothesis.
- We call a learning algorithm *consistent learner* if it returns a hypothesis that commits zero errors over the training examples.

Note that:

Every consistent learner returns an MAP hypothesis, if we assume a uniform prior probability distribution over $\mathcal{H}$ (i.e., $\forall i,j : \Pr(h_i) = \Pr(h_j)$), and if we assume deterministic, noise-free training data (i.e., $\Pr(\mathcal{D}\,|h) = 1$ if $\mathcal{D}$ and h are consistent, 0 otherwise).

Probabilistic Interpretation of Inductive Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Under our assumptions on $\Pr(h)$ and $\Pr(\mathcal{D}\,|h)$, Find-S (Alg. 5.1) provides an MAP hypothesis, although it never manipulates probabilities explicitly.
- Find-S is a consistent learner.
- Find-S returns the MAP hypothesis also for other probability distributions than the fixed values for $\Pr(h)$, if for any hypothesis h_1 that is more specific than h_2, we have $\Pr(h_1) \geq \Pr(h_2)$.
- Identifying probability distributions $\Pr(h)$ and $\Pr(\mathcal{D}\,|h)$ under which some algorithm returns optimal (MAP) hypotheses allows to characterize the implicit assumptions under which that algorithm behaves optimally.
- We can model learning from noisy data by allowing $\Pr(\mathcal{D}\,|h)$ to take on values other than 0 and 1.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

William of Ockham (also: Occam)

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

William of Ockham (~1287–1347): memorial slab in the monastery of the Franciscans in the Lehel, Munich, Germany

Occam's Razor

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

William of Ockham frequently used the principle of parsimony in his philosophical reasoning:

Frustra fit per plura quod potest fieri per pauciora.

(It is pointless to do with more what can be done with fewer.)

Numquam ponenda est pluralitas sine necessitate.

(Plurality is never to be posited without necessity.)

In terms of learning, this can be (debatably) seen as a guiding principle: "choose the shortest explanation for the observed data".

Preference for the Simpler Model

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Remember Kepler (Slide 5):

How the observations look like:

How Kepler's law can be used to describe them:

Encoding of Messages

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Messages for transmission are drawn at random from a pool.
- We encounter message i with probability $\Pr(i)$.
- We are interested in the most compact code (minimizing the number of bits for transmission of a message drawn at random).
- We assign therefore shorter codes to messages that are more probable.
- Optimal code assigns $-\log_2 \Pr(i)$ bits to encode message i [Shannon and Weaver, 1949].
- Number of bits required to encode message i using code C is the *description length of message i w.r.t. C*, denoted by $L_C(i)$.

The Minimum Description Length Principle

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- For $C_{\mathcal{H}}$ being the *optimal* code for hypothesis space $\mathcal{H}$, we have
$$L_{C_{\mathcal{H}}}(h) = -\log_2 \Pr(h)$$
- For $C_{(\mathcal{D}\,|h)}$ being the *optimal* code for describing data $\mathcal{D}$ assuming that both sender and receiver know hypothesis h, we have
$$L_{C_{(\mathcal{D}\,|h)}}(\mathcal{D}\,|h) = -\log_2 \Pr(\mathcal{D}\,|h)$$
- The *minimum description length principle* would recommend to choose
$$h_{\text{MDL}} = \arg\min_{h \in \mathcal{H}} L_{C_{\mathcal{H}}}(h) + L_{C_{(\mathcal{D}\,|h)}}(\mathcal{D}\,|h)$$

Bayesian Perspective on the MDL Principle

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If the encodings $C_{\mathcal{H}}$ and $C_{(\mathcal{D}\,|h)}$ are choosen in the discussed *optimal* way, we have:

$$\begin{aligned} h_{\text{MAP}} &= \arg\max_{h\in\mathcal{H}} \Pr(\mathcal{D}\,|h)\Pr(h) \\ &= \arg\max_{h\in\mathcal{H}} \log_2 \Pr(\mathcal{D}\,|h) + \log_2 \Pr(h) \\ &= \arg\min_{h\in\mathcal{H}} -\log_2 \Pr(\mathcal{D}\,|h) - \log_2 \Pr(h) \\ &= \arg\min_{h\in\mathcal{H}} L_{C_{\mathcal{H}}}(h) + L_{C_{(\mathcal{D}\,|h)}}(\mathcal{D}\,|h) \\ &= h_{\text{MDL}} \end{aligned}$$

In order to show that we actually have such an optimal representation, we need to know all the prior probabilities $\Pr(h)$ as well as $\Pr(\mathcal{D}\,|h)$.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Most Probable Hypothesis vs. Most Probable Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Consider some hypothesis space $\mathcal{H} = \{h_1, h_2, h_3\}$ with $\Pr(h_1|\mathcal{D}) = 0.4$, $\Pr(h_2|\mathcal{D}) = 0.3$, $\Pr(h_3|\mathcal{D}) = 0.3$.
- Obviously, h_1 is the MAP hypothesis.
- Suppose a new instance x is encountered, where

$$h_1(x) = A$$
$$h_2(x) = B$$
$$h_3(x) = B$$

- Taking all hypotheses into account, we have:

$$\Pr(A|x) = 0.4$$
$$\Pr(B|x) = 0.6$$

- The most probable classification is different from the classification generated by the MAP hypothesis.

Bayes Optimal Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We obtain the most probable classification by combining the predictions of all hypotheses weighted by the posterior probabilities.
- For the set of classes C, for any $c_j \in C$, we have

$$\Pr(c_j|\,\mathcal{D}) = \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i)\,\Pr(h_i|\,\mathcal{D})$$

- The optimal classification is therefore:

$$\arg\max_{c_j \in C} \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i)\,\Pr(h_i|\,\mathcal{D})$$

- Any system classifying new instances according to this rule is called a "Bayes optimal classifier".

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

With $C = \{A, B\}$ and the above example, the possible classifications of the new instance x are:

$$\Pr(h_1|\mathcal{D}) = 0.4 \qquad \Pr(A|h_1) = 1 \qquad \Pr(B|h_1) = 0$$
$$\Pr(h_2|\mathcal{D}) = 0.3 \qquad \Pr(A|h_2) = 0 \qquad \Pr(B|h_2) = 1$$
$$\Pr(h_3|\mathcal{D}) = 0.3 \qquad \Pr(A|h_3) = 0 \qquad \Pr(B|h_3) = 1$$

Therefore:

$$\sum_{h_i \in \mathcal{H}} \Pr(A|h_i) \Pr(h_i|\mathcal{D}) = 0.4$$
$$\sum_{h_i \in \mathcal{H}} \Pr(B|h_i) \Pr(h_i|\mathcal{D}) = 0.6$$

and

$$\arg\max_{c_j \in \{A,B\}} \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i) \Pr(h_i|\mathcal{D}) = B$$

Properties of the Bayes Optimal Classifier

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The predictions of the Bayes optimal classifier can correspond to the predictions of a hypothesis that is not contained in the original hypothesis space $\mathcal{H}$!
- The Bayes optimal classifier considers effectively a different hypothesis space $\mathcal{H}'$, including hypotheses that perform comparisons between linear combinations of predictions from multiple hypotheses in $\mathcal{H}$.

Note that:

The Bayes optimal learner maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses. Thus no other classification method using the same hypothesis space and same prior knowledge can outperform this method on average.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - The Minimum Description Length Principle
  - Bayes Optimal Classifier
  - Naïve Bayes Classifier
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Estimates of Prior Probabilities

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Consider a learning task to distinguish apples, oranges, and other fruits, where the objects are described by color and shape:

- 20% of the objects are apples
- 30% of the objects are oranges

} prior class probability

- 50% of the objects are round
- 40% of the objects have an orange color

} prior probability of some attribute value

Estimates of Posterior Probabilities

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Posterior (conditional) probabilities model relations between attribute values and classes:

- 100% of the oranges are round: $\Pr(\text{shape=round}|\text{orange})$
- 100% of the apples are round: $\Pr(\text{shape=round}|\text{apple})$
- 90% of the oranges have the color orange: $\Pr(\text{color=orange}|\text{orange})$

Bayes Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For a given attribute value a_q, we can compute the posterior class probability according to Bayes' rule:

$$\Pr(c_j|a_q) = \frac{\Pr(a_q|c_j)\Pr(c_j)}{\Pr(a_q)} = \frac{\Pr(a_q|c_j)\Pr(c_j)}{\sum_{c_i \in C}\Pr(c_i)\Pr(a_q|c_i)}$$

We estimate probabilities from the training data.
For example, we have an object with color orange:

$$\begin{aligned}
&\Pr(\text{orange}|\text{color=orange}) \\
&\quad = \frac{\Pr(\text{color=orange}|\text{orange})\Pr(\text{orange})}{\Pr(\text{color=orange})} \\
&\quad = \frac{0.9 \cdot 0.3}{0.4} \\
&\quad = 0.675
\end{aligned}$$

Maximum Likelihood Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given all posterior class probabilities, we predict the most likely class:

$$\begin{aligned} h_{\mathsf{MAP}} &= \arg\max_{c_i \in C} \Pr(c_i | a_q) \\ &= \arg\max_{c_i \in C} \frac{\Pr(a_q | c_i) \Pr(c_i)}{\Pr(a_q)} \\ &= \arg\max_{c_i \in C} \Pr(a_q | c_i) \Pr(c_i) \end{aligned}$$

Discrete Attributes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

we can count relative frequencies to estimate probabilities:

ID	shape	color	class
1	round	orange	A
2	round	green	A
3	round	yellow	A
4	square	green	A
5	oval	white	B

$$\Pr(\text{shape=round}|A) = \frac{3}{4}$$

$$\Pr(\text{color=green}|A) = \frac{2}{4}$$

$$\Pr(\text{shape=oval}|A) = \frac{0}{4}$$

Continuous Metric Attributes

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\Pr(9.0 < \text{diameter} \leq 9.5 | A) = 10\%$$
$$\Pr(9.5 < \text{diameter} \leq 10.0 | A) = 30\%$$
$$\Pr(10.0 < \text{diameter} \leq 10.5 | A) = 30\%$$
$$\Pr(10.5 < \text{diameter} \leq 11.0 | A) = 10\%$$
$$\Pr(11.0 < \text{diameter} \leq 11.5 | A) = 5\%$$

Zero Probabilities?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Note that:

- *Problem:* $\Pr(\textit{shape=oval}|A) = 0$ *would rule out any slight possibility of predicting an instance of class* A.
- *Heuristic solution: smoothing (use some artificial small minimum probability):*

$$\Pr(\text{shape} = \text{oval}|A) := \max\left\{\frac{0}{4}, \varepsilon\right\} \textit{ with } 0 < \varepsilon \ll 1$$

Multi-dimensional Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- So far, we considered only one attribute.
- In multi-dimensional data, we need to estimate the combined probabilities of specific attribute values:

$$\begin{aligned} h_{\text{MAP}} &= \arg\max_{c_i \in C} \Pr(c_i | a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n) \\ &= \arg\max_{c_i \in C} \frac{\Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i)}{\Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n)} \\ &= \arg\max_{c_i \in C} \Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i) \end{aligned}$$

Example:

ID	*shape*	*color*	*class*
1	*round*	*orange*	*A*
2	*round*	*green*	*A*
3	*round*	*yellow*	*A*
4	*square*	*green*	*A*
5	*oval*	*white*	*B*

$$\Pr(\textit{shape=round} \cap \textit{color=orange} | A) = \frac{1}{4}$$

$$\Pr(\textit{shape=round} \cap \textit{color=green} | A) = \frac{1}{4}$$

Problems for the Bayes Classifier in Multi-dimensional Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Problems:

- If we have n attributes, and each can take on r different values, we have r^n different attribute combinations.
- Typically, there are not enough training instances available to reliably estimate probabilities.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

ID	shape	color	class
1	round	orange	A
2	round	green	A
3	round	yellow	A
4	square	green	A
5	oval	white	B

$$\Pr(\text{shape=round} \cap \text{color=orange}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=green}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=yellow}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=orange}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=green}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=yellow}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=orange}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=green}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=yellow}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=orange}|B) = \frac{0}{1}$$

$$\vdots$$

The probability estimates are unreliable, because the sample size is too small for each instance.

The Naïve Assumption: Independence

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If we assume independence among the attributes, we can estimate the combined probabilities based on Definition 6.5 ($\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$):

$$h_{\text{MAP}} = \arg\max_{c_i \in C} \Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i)$$

$$= \arg\max_{c_i \in C} \prod_{j=1}^{n} \Pr(a_j | c_i) \Pr(c_i) \qquad \text{(Ass. of Indep.)}$$

- The assumption might be wrong.
- Then we don't get the correct probabilities.
- But we *might* still get the correct maximum.
- In practice, the assumption often works despite *some* dependency among the attributes.

Naïve Bayes Classifier

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The "Naïve Bayes classifier" relies on the assumption of independence of attributes.
- The various $\Pr(c_i)$ and $\Pr(a_j|c_i)$ terms are estimated based on the relative frequencies of corresponding examples in the training data.
- The set of these estimates constitutes the learned hypothesis.
- The class prediction is based on these estimates according to:

$$h_{\text{naïve Bayes}} = \arg\max_{c_i \in C} \prod_{j=1}^{n} \Pr(a_j|c_i) \Pr(c_i)$$

- Because of the multiplication, the replacement of zero probabilities by some heuristic minimum probability ε (cf. slide 316) is particularly important.

Example: Should We Play Tennis Today?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

ID	forecast	temperature	humidity	wind	play tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example Prediction

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Classify a new instance: $\langle \text{sunny}, \text{cool}, \text{high}, \text{strong} \rangle$

$$h_{\text{naïve Bayes}} = \arg\max_{c_i \in \{\text{yes,no}\}} \prod_{j=1}^{n} \Pr(a_j|c_i)\Pr(c_i)$$

$$= \arg\max_{c_i \in \{\text{yes,no}\}} \Pr(\text{sunny}|c_i)\Pr(\text{cool}|c_i)\Pr(\text{high}|c_i)$$

$$\Pr(\text{strong}|c_i)\Pr(c_i)$$

$$\Pr(\text{yes}) = \frac{9}{14} = 0.64 \qquad \Pr(\text{wind=strong}|\text{yes}) = \frac{3}{9} = 0.33$$

$$\Pr(\text{no}) = \frac{5}{14} = 0.36 \qquad \Pr(\text{wind=strong}|\text{no}) = \frac{3}{5} = 0.60$$

$$\vdots$$

Decision and Probability

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The Naïve Bayes classifier decides by finding the class maximizing the product of probabilities:

$$\Pr(\text{sunny}|\text{yes})\Pr(\text{cool}|\text{yes})\Pr(\text{high}|\text{yes})\Pr(\text{strong}|\text{yes})\Pr(\text{yes}) = 0.0053$$
$$\Pr(\text{sunny}|\text{no})\Pr(\text{cool}|\text{no})\Pr(\text{high}|\text{no})\Pr(\text{strong}|\text{no})\Pr(\text{no}) = 0.0206$$

$$h_{\text{naïve Bayes}}(\langle\text{sunny}, \text{cool}, \text{high}, \text{strong}\rangle) = \text{no}$$

If we are interested in the conditional probability for "no", we could normalize these quantities to sum up to one:

$$\frac{0.0206}{0.0206 + 0.0053} = 0.795$$

Assumption of Independence is a Bias

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
MDL
Bayes Optimal
Naïve Bayes
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The assumption of independence can be seen as the bias inherent to the Naïve Bayes classifier.
- An unbiased probabilistic classifier is not practical due to a notorious lack of training examples.
  - In other words: in any practical scenario, it would hopelessly overfit.
- For data consisting of two classes and only 30 binary attributes, we would need more than 2 billion examples just to see each combination *once* (which is not good enough to derive reliable probability estimates).
- Relying on the bias, the classifier may have a tendency to be wrong (if the assumption does not hold).
- The bias is necessary to make generalization feasible.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *Axioms of probability:*
  - *sample space*
  - *event*
  - *probability function*
  - *probability space*
- *independence and conditional probability*
- *probabilistic interpretation of quality measures for association rules*
- *Bayes' rule*

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *Bayesian Learning:*
  - *k nearest neighbor classifier as an application of Bayes' rule for learning*
  - *The principle of Bayesian learning:*
    - *prior and posterior probabilities*
    - *data as evidence to adapt probability estimates and to select hypotheses*
    - *MAP hypothesis*
    - *the MDL principle and Bayesian reasoning*
  - *Bayes optimal classifier*
  - *Naïve Bayes classifier*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
  - Random Variables and Expectation
  - The Bernoulli and Binomial Random Variables
  - Conditional Expectation
  - The Geometric Distribution
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Mitzenmacher and Upfal [2017], Chapter 2.*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
  - Random Variables and Expectation
  - The Bernoulli and Binomial Random Variables
  - Conditional Expectation
  - The Geometric Distribution
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Event vs. Values Associated with the Event

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- When throwing two dice, we have the 36 basic events of equal probability given by the ordered pair of numbers $\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$.
- As opposed to being interested in some basic event as such, we are often interested in some more complex event, e.g., the sum of the two dice.
- We have in this case actually 11 different events (the possible outcomes of the sum).
- These 11 events have unequal probability.
- We can map any of the basic events to some of the 11 complex events (sum is 2, 3, . . . , 12).
- Any such function from the sample space to the real numbers is called a *random variable*.

Random Variable

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.1 (Random Variable)

A *random variable* X on a sample space Ω is a real-valued function from Ω to the real numbers:

$$X : \Omega \to \mathbb{R}$$

Definition 7.2 (Discrete Random Variable)

A *discrete random variable* is a random variable that takes on only a finite or countably infinite number of values.

For example, for some $I \subseteq \mathbb{R}$:

$$X : \Omega \to I$$

is a *discrete random variable*, if I is finite or countably infinite (e.g., $I = \mathbb{N}$, $I = \{0, 1\}$).

Random Variables and Basic Events

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- For a discrete random variable X and $a \in \mathbb{R}$, the event "$X = a$" includes all basic events of the sample space in which the random variable X assumes the value a.
- "$X = a$" represents the set $\{s \in \Omega | X(s) = a\}$.
- The probability of that event is given by:

$$\Pr(X = a) = \sum_{s \in \Omega : X(s) = a} \Pr(s)$$

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Let X be the random variable representing the sum of two dice.
- $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(6,6)\}$
- The event $X = 4$ corresponds to the set of basic events: $\{(1,3),(2,2),(3,1)\}$:

$$X((1,3)) = 4; \quad X((2,2)) = 4; \quad X((3,1)) = 4;$$

$$\forall (x,y) \in \Omega \setminus \{(1,3),(2,2),(3,1)\} : X((x,y)) \neq 4$$

- Therefore:

$$\Pr(X = 4) = \frac{3}{36} = \frac{1}{12}$$

Probability Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.3 (Probability Distribution of a Discrete Random Variable)

The *probability distribution* (or: *probability (mass) function (pmf)*) of a discrete random variable X is defined for every value x in the range of X by:

$$p_X(x) = \Pr(X = x) = \sum_{s \in \Omega, X(s) = x} \Pr(s),$$

where $p_X(x) \geq 0$ and $\sum_{-\infty < x < \infty} \Pr(x) = 1$.

The probability distribution of X says how the total probability of 1 is distributed among all possible values of X.

Probability Distribution: Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(cf. slide 315:)

$$\Pr(9.0 < \text{diameter} \leq 9.5|A) = 10\%$$
$$\Pr(9.5 < \text{diameter} \leq 10.0|A) = 30\%$$
$$\Pr(10.0 < \text{diameter} \leq 10.5|A) = 30\%$$
$$\Pr(10.5 < \text{diameter} \leq 11.0|A) = 10\%$$
$$\Pr(11.0 < \text{diameter} \leq 11.5|A) = 5\%$$

Cumulative Distribution Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.4 (Cumulative Distribution Function of a Discrete Random Variable)

The *cumulative distribution function (cmf)* of a discrete random variable X with pmf $p(x)$ is defined for every value x by:

$$F(x) = \Pr(X \leq x) = \sum_{y:y \leq x} p(y).$$

For any number x, the value of the cmf $F(x)$ is the probability that the observed value of X will be at most x.

Cumulative Distribution Function: Example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\Pr(\text{diameter} \leq 9.5|A) = 10\%$$
$$\Pr(\text{diameter} \leq 10.0|A) = 40\%$$
$$\Pr(\text{diameter} \leq 10.5|A) = 70\%$$
$$\Pr(\text{diameter} \leq 11.0|A) = 80\%$$
$$\Pr(\text{diameter} \leq 11.5|A) = 85\%$$

(cf. slide 315)

Independent Events

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The definition of independence of events extends to random variables:

Definition 7.5 (Independence of Random Variables)

Two random variables X and Y are independent if and only if

$$\Pr((X = x) \cap (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y)$$

for all values x and y.

Similarly, random variables $X_1, X_2, \ldots, X_k$ are mutually independent if and only if, for any subset $I \subseteq [1, k]$ and any values $x_i, i \in I$:

$$\Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i)$$

Expectation (or: Mean)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The *expectation* (or: *mean*) is a basic characteristic of a random variable.
- It is the weighted average of the values assumed by the random variable.
- The weight of a value is the probability that the variable assumes that value.

Definition 7.6 (Expectation)

The *expectation* of a discrete random variable X is given by:

$$E[X] = \mu_X = \sum_i i \Pr(X = i),$$

where the summation is over all values i in the range of X. The expectation is finite if $\sum_i |i| \Pr(X = i)$ converges; otherwise, the expectation is unbounded.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The expectation of the random variable X representing the sum of two dice is:

$$
\begin{aligned}
E[X] = \quad & 2 \cdot \frac{|\{(1,1)\}|}{|\Omega|} \\
& + 3 \cdot \frac{|\{(1,2),(2,1)\}|}{|\Omega|} \\
& + 4 \cdot \frac{|\{(1,3),(2,2),(3,1)\}|}{|\Omega|} \\
& + \ldots \\
& + 12 \cdot \frac{|\{(6,6)\}|}{|\Omega|} \\
= \quad & 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \ldots + 12 \cdot \frac{1}{36} \\
= \quad & 7
\end{aligned}
$$

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The expectation of the random variable X representing the value of a single die is:

$$\begin{aligned} E[X] &= 1 \cdot \frac{|\{1\}|}{|\Omega|} + 2 \cdot \frac{|\{2\}|}{|\Omega|} + 3 \cdot \frac{|\{3\}|}{|\Omega|} + \ldots + 6 \cdot \frac{|\{6\}|}{|\Omega|} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} \\ &= \frac{21}{6} \\ &= 3.5 \end{aligned}$$

Note that:

The expected value of X, $E(X)$, is not necessarily a possible value of X.

Linearity of Expectations

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.7 (Linearity of Expectations)

For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$ with finite expextations,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Proof: see Mitzenmacher and Upfal [2017], p. 25.

Lemma 7.8

For any constant c and discrete random variable X:

$$E[cX] = cE[X]$$

Proof: see Mitzenmacher and Upfal [2017], p. 26.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Let $X = X_1 + X_2$ be the sum of two standard dice, where $X_i, i = 1, 2$ represents the outcome of die i.

$$E[X_i] = \sum_{j=1}^{6} \frac{1}{6} \cdot j = \frac{7}{2}$$

$$E[X] = E[X_1] + E[X_2] = 7$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
  - Random Variables and Expectation
  - The Bernoulli and Binomial Random Variables
  - Conditional Expectation
  - The Geometric Distribution
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Random Variable with Parameter

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.9 (Bernoulli Random Variable)

A random variable $Y : \Omega \to \{0, 1\}$ is a *Bernoulli* (or: *indicator*) random variable.
A *Bernoulli* random variable Y has one parameter $0 \leq p \leq 1$:

$$Y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

Note that:

If an experiment succeeds with probability p and fails with probability $1 - p$, the expected value for a corresponding Bernoulli random variable equals the winning probability:

$$E[Y] = 1 \cdot p + 0 \cdot (1 - p) = p = \Pr(Y = 1)$$

Examples

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We flip a fair coin and call the outcome "head" a success.
  - The expected outcome of the corresponding random variable is $\frac{1}{2}$.
- In a game, we throw two dice and win if (and only if) both dice land on 6.
  - The expected outcome of the random variable is $\frac{1}{36}$.
- N people checked coats in a restaurant. The coats are mixed and each person gets a random coat when checking out. What is the probability for a given person to get her own coat?
  - Define N random variables X_i, where $X_i = 1$ if and only if person i got her coat.
  - $\Pr(X_i = 1) = \frac{1}{N}$

Binomial Random Variable

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- To model a sequence of n independent experiments, where each succeeds with probability p, we define a random variable with two parameters.
- Let X represent the number of successes in the n experiments.
- X has a binomial distribution.

Definition 7.10 (Binomial Random Variable)

A *binomial random variable* X with parameters n and p, denoted by $B(n,p)$, is defined by the following probability distribution on $j = 0, 1, 2, \ldots, n$:

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}$$

Expectation of a Binomial Random Variable

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Proposition 7.11

The expectation $E(X)$ of a binomial random variable $X = B(n,p)$ is given by $n \cdot p$.

Proof.

$$E[X] = \sum_{j=0}^{n} j \binom{n}{j} p^j (1-p)^{n-j} = \sum_{j=0}^{n} j \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j} = \sum_{j=1}^{n} \frac{n!}{(j-1)!(n-j)!} p^j (1-p)^{n-j}$$

$$= np \sum_{j=1}^{n} \frac{(n-1)!}{(j-1)!((n-1)-(j-1))!} p^{j-1} (1-p)^{(n-1)-(j-1)}$$

$$= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!((n-1)-k)!} p^k (1-p)^{(n-1)-k} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k}$$

$= np$ (The binomial identity: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$ is used in the last equation.)

$\square$

Expectation of a Binomial Random Variable

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Simpler proof by Theorem 7.7.

- If X is a binomial random variable with parameters n and p, then X is the number of successes in n trials, where each trial is successful with probability p.
- Define a set of n indicator random variables $X_1, X_2, \ldots, X_n$, where $X_i = 1$ if the ith trial is successful and 0 otherwise.
- $E[X_i] = p$
- $X = \sum_{i=1}^{n} X_i$

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np$$

$\square$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
  - Random Variables and Expectation
  - The Bernoulli and Binomial Random Variables
  - Conditional Expectation
  - The Geometric Distribution
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Conditional Expectation of a Random Variable

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.12 (Conditional Expectation)

The conditional expectation of a random variable Y, given another random variable Z assumes the value z, is defined as:

$$E[Y|Z=z] \equiv \sum_y y \Pr(Y=y|Z=z),$$

where the sum is over all y in the range of Y.

- Like the expectation, also the conditional expectation of a random variable is a weighted sum of the values it assumes.
- The weights for the conditional expectation are given by the *conditional probability* that the variable assumes that value.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Suppose that we independently roll two dice. Let X_1 be the number that shows on the first die, X_2 the number on the second die, and X the sum of the numbers on the two dice.

$$E[X|X_1 = 2] = \sum_x x \Pr(X = x|X_1 = 2) = \sum_{x=3}^{8} x \cdot \frac{1}{6} = \frac{11}{2}$$

$$\begin{aligned} E[X_1|X = 5] &= \sum_{x=1}^{4} x \Pr(X_1 = x|X = 5) \\ &= \sum_{x=1}^{4} x \frac{\Pr(X_1 = x \cap X = 5)}{\Pr(X = 5)} \\ &= \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2} \end{aligned}$$

Extension of Total Probability to Expectations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 7.13

For any random variables X and Y:

$$E[X] = \sum_y \Pr(Y = y) E[X|Y = y],$$

where the sum is over all values in the range of Y and all of the expectations exist.

Cf. Theorem 6.7.

Extension of Linearity of Expectations to Conditional Expectations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 7.14

For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$ with finite expectations and for any random variable Y,

$$E\left[\sum_{i=1}^{n} X_i | Y = y\right] = \sum_{i=1}^{n} E[X_i | Y = y].$$

The Conditional Expectation *is* a Random Variable

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.15

The expression $E[Y|Z]$ is a random variable $f(Z)$ that takes on the value $E[Y|Z = z]$ when $Z = z$.

Note that:

$E[Y|Z]$ is not a real value; rather it is a function of the random variable Z.
Thus $E[Y|Z]$ is itself a function from the sample space to the real numbers and can be thought of as a random variable.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recurring to the example rolling two dice (Slide 355):

$$E[X|X_1] = \sum_x x \Pr(X = x|X_1) = \sum_{x=X_1+1}^{X_1+6} x \cdot \frac{1}{6} = X_1 + \frac{7}{2}$$

- Obviously, $E[X|X_1]$ is a random variable whose value depends on X_1.
- We can therefore also consider its expectation:

$$E[E[X|X_1]] = E\left[X_1 + \frac{7}{2}\right] = \frac{7}{2} + \frac{7}{2} = 7 = E[X].$$

(Cf. Slide 346.)

Expectation of $E[Y|Z]$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.16

$$E[Y] = E[E[Y|Z]].$$

Proof.

From Def. 7.15 we have $E[Y|Z] = f(Z)$, where $f(Z)$ takes on the value $E[Y|Z = z]$ when $Z = z$. Hence, by Lemma 7.13:

$$E[E[Y|Z]] = \sum_z E[Y|Z = z] \Pr(Z = z) = E[Y]$$

□

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Discrete Random Variables and Expectation
Random Variables and Expectation
The Bernoulli and Binomial Random Variables
Conditional Expectation
The Geometric Distribution
Moments and Deviations
Continuous Distributions
Bayesian Learning with Distributions (Parametric Learning)
Non-Parametric Learning
Outlier Detection — Parametric and Non-Parametric
Summary
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Geometric Random Variable

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.17 (Geometric Random Variable)

A *geometric random variable* X with parameter p $(0 \leq p \leq 1)$ is given by the following probability distribution on $n = 1, 2, \ldots$:

$$\Pr(X = n) = (1 - p)^{n-1} p$$

The geometric random variable describes the number of failures before the first success: for $X = n$, we must have $n - 1$ failures, followed by a success.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

N people checked coats in a restaurant. The coats are mixed and each person gets a random coat when checking out. How many coats has a person to try before getting her own?

- Recall that each individual trial follows a Bernoulli distribution (example on slide 349).
- The series of random trials follows a geometric distribution with parameter $p = \frac{1}{N}$.

Geometric Random Variables are Memoryless

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

We call geometric random variables *memoryless* because the probability of reaching the first success after n trials *from now on* is independent of the number of failures experienced in the past.

Lemma 7.18 (Memoryless Property)

For a geometric random variable X with parameter p and for $n > 0$,

$$\Pr(X = n + k | X > k) = \Pr(X = n)$$

- Proof: Mitzenmacher and Upfal [2017], p. 33.
- Intuition:
  - I toss a coin and hand it to you. You toss the same coin.
  - The probability that you get the first head after three trials is independent of how often I tossed the same coin before and failed to get head.

Expectation of Random Variables $\in \Omega \to \mathbb{N}_0$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 7.19

Let X be a discrete random variable that takes on only non-negative integer values. Then

$$E[X] = \sum_{i=1}^{\infty} \Pr(X \geq i).$$

- Proof: Mitzenmacher and Upfal [2017], p. 34.
- Verify:

Expectation of a Geometric Random Variable

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
RV and Exp.
Bernoulli and Binomial RV
Conditional Exp.
Geom. Distr.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Recall (from mathematical analysis): $\sum_{i=k}^{\infty} x^i = \frac{x^k}{1-x}$.
- For a geometric random variable X with parameter p,

$$\Pr(X \geq i) = \sum_{n=i}^{\infty} (1-p)^{n-1} p = (1-p)^{i-1}$$

- Due to Lemma 7.19, we have therefore:

$$E[X] = \sum_{i=1}^{\infty} \Pr(X \geq i) = \sum_{i=1}^{\infty} (1-p)^{i-1}$$
$$= \frac{1}{1-(1-p)} = \frac{1}{p}$$

- On average, with two flips of a fair coin you get the first head.
- On average, the guests try N coats.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
  - Markov's Inequality
  - Variance and Moments of a Random Variable
  - Chebyshev's Inequality
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Mitzenmacher and Upfal [2017], Chapter 3.*
- *Not strictly technical reading but nevertheless quite entertaining are the books by Taleb [2001, 2007].*

Expectation and Risk

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- I propose a game: With probability 0.99, you pay me 10 kr., with probability 0.01, I pay you 1000 kr.

- Would you play this game?

Expectation and Risk

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

With probability 0.99, you pay me 10 kr., with probability 0.01, I pay you 1000 kr.

- $X =$ your winning in kr.
- $Y =$ my winning in kr.

Expectation and Risk

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

With probability 0.99, you pay me 10 kr., with probability 0.01, I pay you 1000 kr.

- $X =$ your winning in kr.
- $Y =$ my winning in kr.
- $E[X] = 0.1$
- $E[Y] = -0.1$

Expectation and Risk

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

With probability 0.99, you pay me 10 kr., with probability 0.01, I pay you 1000 kr.

- $X =$ your winning in kr.
- $Y =$ my winning in kr.
- $E[X] = 0.1$
- $E[Y] = -0.1$
- $\Pr(X \geq E[X]) = 0.01$ (actually, $\Pr(X \geq 0) = 0.01$)
- $\Pr(Y \geq E[Y]) = 0.99$ (actually, $\Pr(Y > 0) = 0.99$)

Expectation and Risk

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You need to select (and buy) a program for a common task in your company:

- The expected run time of program A is 1 hour but it will take 100 hours in 1% of the runs.
- The run time of program B is always 2 hours.
- Which program do you prefer?

We want to bound the probability that the run time of some program deviates significantly from its average.

More general: we want to bound the *tail distribution*, the probability that a random variable assumes values that are far from its expectation.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
  - Markov's Inequality
  - Variance and Moments of a Random Variable
  - Chebyshev's Inequality
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Markov's Inequality

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.20 (Markov's Inequality)

Let X be a random variable that assumes only non-negative values. Then, for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{E[X]}{a}.$$

Markov's Inequality – Proof

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Proof.

For $a > 0$, let $I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$

Since $X \geq 0$,

$$I \leq \frac{X}{a} \tag{7.1}$$

As I is a Bernoulli random variable, we have (see Def. 7.9) $E[I] = \Pr(I = 1) = \Pr(X \geq a)$.
Thus, taking expectations in Eq. 7.1 yields:

$$\Pr(X \geq a) = E[I] \leq E\left[\frac{X}{a}\right] = \frac{E[X]}{a}$$

□

Markov's Inequality – Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

What is the probability of obtaining at least $\frac{3n}{4}$ heads in a sequence of n fair coin flips?

- Let $X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is heads} \\ 0 & \text{otherwise} \end{cases}$
- Let $X = \sum_{i=1}^{n} X_i$ denote the number of heads in the n coin flips.
- Since $E[X_i] = \Pr(X_i = 1) = \frac{1}{2}$, we have

$$E[X] = \sum_{i=1}^{n} E[X_i] = \frac{n}{2}$$

- Applying Markov's inequality, we obtain:

$$\Pr\left(X \geq \frac{3n}{4}\right) \leq \frac{E[X]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}$$

Best Bound for Limited Information

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Note that:

Markov's inequality is the strongest possible bound on the tail distribution when we only know the expectation of the random variable and that it is non-negative.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
  - Markov's Inequality
  - Variance and Moments of a Random Variable
  - Chebyshev's Inequality
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Moments of a Random Variable

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.21 (Moments of a Random Variable)

The kth *moment* of a random variable X is $E[X^k]$.

Note that:

Obviously, the expectation of a random variable can also be addressed as its first moment.

Variance and Standard Deviation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.22 (Variance)

The *variance* of a random variable X is given by:

$$\mathrm{Var}[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - (E[X])^2$$

In retrospect, understand the name of Algorithm 4.1.

Definition 7.23 (Standard Deviation)

The *standard deviation* of a random variable X is

$$\sigma[X] = \sqrt{\mathrm{Var}[X]}$$

Note that:

If a random variable is constant, then its variance and standard deviation are zero.

Covariance

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.24 (Covariance)

The *covariance* of two random variables X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Theorem 7.25

For any two random variables X and Y,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}(X, Y)$$

Proof: see Mitzenmacher and Upfal [2017], p. 49

Correlation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.26 (Correlation)

The *correlation coefficient* of two random variables X and Y is given by

$$\mathrm{Corr}(X,Y) = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}$$

Independence and Expectation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.27

If X and Y are two independent random variables, then

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Corollary 7.28

If X and Y are independent random variables, then

$$\mathrm{Cov}(X, Y) = 0$$

and

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Proofs: see Mitzenmacher and Upfal [2017] pp. 49f.

Variance of Sum and Sum of Variances

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.29

Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables. Then

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$$

Proof: by induction from Corrolary 7.28.

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

We flip a coin and get a point for head, zero points for tail.

- $\Pr(X = 0) = \Pr(X = 1) = \frac{1}{2}$
- $E[X] = \frac{1}{2}$

$$\begin{aligned}\mathrm{Var}[X] &= E\left[(X - E[X])^2\right] \\ &= \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 \\ &= \frac{1}{4} \\ \sigma[X] &= \frac{1}{2}\end{aligned}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
  - Markov's Inequality
  - Variance and Moments of a Random Variable
  - Chebyshev's Inequality
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Chebyshev's Inequality

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If we know expectation *and* variance of some random variable, we can derive stronger bounds on the tail distribution:

Theorem 7.30 (Chebyshev's Inequality)

For any $a > 0$*:* $$\Pr(|X - E[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$$

Proof.

$$\begin{aligned}\Pr(|X - E[X]| \geq a) &= \Pr\left((X - E[X])^2 \geq a^2\right) \\ &\leq \frac{E\left[(X - E[X])^2\right]}{a^2} && \text{(by Theorem 7.20)} \\ &= \frac{\mathrm{Var}[X]}{a^2}\end{aligned}$$

□

Chebyshev's Inequality – Variant

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Corollary 7.31

For any $z > 1$,

$$\Pr(|X - E[X]| \geq z \cdot \sigma[X]) \leq \frac{1}{z^2}$$

Chebyshev's Inequality – Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

What is the probability of obtaining at least $\frac{3n}{4}$ heads in a sequence of n fair coin flips (cf. Slide 375)?

- $X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is heads} \\ 0 & \text{otherwise} \end{cases}$
- $X = \sum_{i=1}^{n} X_i$: number of heads in the n coin flips.
- $E[X_i] = \Pr(X_i = 1) = \frac{1}{2}$, $\mathrm{Var}[X_i] = \frac{1}{4}$
- $E[X] = \sum_{i=1}^{n} E[X_i] = \frac{n}{2}$
- $\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{n}{4}$ (by Theorem 7.29)

$$\Pr\left(X \geq \frac{3n}{4}\right) \leq \Pr\left(|X - E[X]| \geq \frac{n}{4}\right) \leq \frac{\mathrm{Var}[X]}{(n/4)^2} = \frac{n/4}{(n/4)^2} = \frac{4}{n}$$

By symmetry:

$$\Pr\left(X \geq \frac{3n}{4}\right) \leq \frac{2}{n}$$

Application: Advantage of Multiple Samples

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Lemma 7.32

For any random variable X and constant a:
$\mathrm{Var}[aX] = a^2 \,\mathrm{Var}[X]$.

Lemma 7.33

Let $X_1, \ldots, X_n$ be n independent, identically distributed random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\,\mathrm{Var}[X_i]$$

Application: Advantage of Multiple Samples

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.34 (The Weak Law of Large Numbers)

Let $X_1, \ldots, X_n$ be n independent, identically distributed random variables with mean μ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| > \varepsilon\right) = 0$$

Mean and Median

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

For any random variable X with finite expectation $E[X]$ and finite median m,

1. the expectation $E[X]$ (= mean) is the value of c that minimizes $E[(X-c)^2]$
2. the median m is a value of c that minimizes $E[|X-c|]$.

Note that:

This is why we use different objective functions, TD^2 and TD, for k-means and k-medians, respectively (see Slides 157 and 170).

Outlook: Chernoff Bounds

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Markov's Inequality
Variance and Moments of a RV
Chebyshev's Inequality
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We tried to show how unlikely it is that a random variable X assumes values far from its expectation.
- The more is known about X, the tighter bounds we can give for the tail distribution:
  - Markov's inequality uses $E[X]$.
  - Chebyshev's inequality uses $\mathrm{Var}[X]$.
  - Chernoff bounds use higher moments by means of *moment generating functions*.

Recommended Reading:

For Chernoff bounds see Mitzenmacher and Upfal [2017], Chapter 4. (But that material is beyond the scope of this lecture.)

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
  - Continuous Random Variables
  - The Uniform Distribution
  - The Normal Distribution
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *The material of this section is covered by Mitzenmacher and Upfal [2017], Chapters 8&9.*
- *Broad coverage (although following a different systematic) is given by Devore and Berk [2012], scattered over Chapters 4&5.*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
  - Continuous Random Variables
  - The Uniform Distribution
  - The Normal Distribution
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Probability Distributions in $\mathbb{R}$

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A continuous roulette wheel has circumference 1.
- The outcome when the wheel stops after spinning is the clockwise distance x from the "0" to the arrow.

- $\Omega = [0, 1)$
- Any point on the circumference of the disk is equally likely to face the arrow when the disk stops.
- What is the probability p of a given outcome x?

Note that:

Notation for subsets of $\mathbb{R}$:

$$[a, b] = \{x | x \in \mathbb{R}, a \leq x \leq b\} \quad [a, b) = \{x | x \in \mathbb{R}, a \leq x < b\}$$
$$(a, b] = \{x | x \in \mathbb{R}, a < x \leq b\} \quad (a, b) = \{x | x \in \mathbb{R}, a < x < b\}$$

Probabilities in an Infinite Sample Space

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recall Definition 6.1:

1. $\forall E : 0 \leq \Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$; and
3. for any finite or countably infinite sequence of pairwise mutually disjoint events E_i: $\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr(E_i)$.

- Let $S(k)$ be a set of k distinct points in the range $[0, 1)$, let p be the probability that any given point in $[0, 1)$ is the outcome of the experiment.
- Since the probability of any event is bounded by 1, we have: $\Pr(x \in S(k)) = kp \leq 1)$
- Since we can choose any integer k, to ensure $kp \leq 1$, we must have $p = 0$.

Probabilities in an Infinite Sample Space

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Note that:

In an infinite sample space

- *there may be possible events that have probability 0;*
- *there may be events with probability 1 (e.g., the complement of an event with probability 0) that do not correspond to all possible outcomes (and therefore are not certain).*

Cumulative Distribution Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.35 (Cumulative Distribution Function)

The probability distribution of a random variable X is given by its *(cumulative) distribution function* $F(x)$, where for any $x \in \mathbb{R}$:

$$F(x) = \Pr(X \leq x).$$

We say that a random variable X is *continuous*, if its distribution function $F(x)$ is a continuous function of x.

Note that:

- *For a continuous random variable X, we must have $\Pr(X = x) = 0$ for any specific value of x.*
- *As a consequence, we have $\Pr(X \leq x) = \Pr(X < x)$*

Probability Density Function

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.36 (Probability Density Function)

If there is a function $f(x)$ such that, for all $-\infty < a < \infty$,

$$F(a) = \int_{-\infty}^{a} f(t)dt,$$

then $f(x)$ is called the *(probability) density function* of $F(x)$. Where the derivative of F is well-defined, we have

$$f(x) = F'(x).$$

For events in $\mathbb{R}$, the integral is the analogue to the sum of probabilities of simple events (cf. Def. 6.2):

$$\Pr(a \leq X < b) = \Pr(X \in [a, b)) = \int_{a}^{b} f(x)dx$$

Joint Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.37 (Joint Distribution)

The *joint distribution function* of X and Y is

$$F(x, y) = \Pr(X \leq x, Y \leq y).$$

The variables X and Y have a joint density function f if, for all x, y:

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) du dv$$

When the derivative exists, we denote

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
  - Continuous Random Variables
  - The Uniform Distribution
  - The Normal Distribution
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The Uniform Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.38 (Uniform Distribution)

When a random variable X assumes values in the interval $[a, b]$ such that all subintervals of equal length have equal probability, we say that

- X has a *uniform distribution* over the interval $[a, b]$
- X is uniform over the interval $[a, b]$

The uniform distribution is denoted by $\mathcal{U}[a, b]$.
The random variable X following a uniform distribution is denoted by $X \sim \mathcal{U}[a, b]$.

Note that:

Since the probability of taking on any specific value when $b > a$ is 0, the distributions $\mathcal{U}[a, b]$, $\mathcal{U}[a, b)$, $\mathcal{U}(a, b]$, and $\mathcal{U}(a, b)$ are essentially the same.

Probability Distribution and Density Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The probability distribution function of a random variable X that is uniform over $[a, b]$ is:

$$F(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } x \geq b \end{cases}$$

and its density function is

$$f(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{if } x > b \end{cases}$$

$$E[X] = \frac{b+a}{2}, \qquad \mathrm{Var}[X] = \frac{(b-a)^2}{12}$$

Example: $\mathcal{U}[0, 1]$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$f(x)$:

$F(x)$:

Two-dimensional Uniform Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

As an example of a joint distribution of two in the interval $[0, 1]$ independently uniformly distributed random variables X and Y, we have a two-dimensional uniform distribution with density function

$$f(x, y) = \begin{cases} 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{otherwise} \end{cases}$$

Two-dimensional Uniform Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The joint probability distribution of $X \sim \mathcal{U}[0,1]$ and $Y \sim \mathcal{U}[0,1]$ is

$$F(x,y) = \begin{cases} 0 & \text{if } x < 0 \vee y < 0 \\ xy & \text{if } (x,y) \in [0,1]^2 \\ x & \text{if } 0 \leq x \leq 1 \wedge y > 1 \\ y & \text{if } 0 \leq y \leq 1 \wedge x > 1 \\ 1 & \text{if } x > 1 \wedge y > 1 \end{cases}$$

or equivalently

$$F(x,y) = \Pr(X \leq x, Y \leq y) = \int_0^{\min\{1,x\}} \int_0^{\min\{1,y\}} 1 \, dy \, dx$$

Two-dimensional Uniform Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

More general: let $S \subset \mathbb{R}^2$ such that the area of S is finite.

$$f(x,y) = \begin{cases} \frac{1}{\text{area}(S)} & \text{if } (x,y) \in S \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr((X,Y) \in B) = \frac{\text{area}(B \cap S)}{\text{area}(S)}$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
  - Continuous Random Variables
  - The Uniform Distribution
  - The Normal Distribution
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The Normal Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.39 (Normal Distribution)

A random variable X follows a *normal distribution* (also: *Gaussian*) with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, if the pdf of X for $-\infty < x < \infty$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The normal distribution is denoted as $\mathcal{N}(\mu, \sigma^2)$.

Note that:

For the parameters μ and σ^2 we have:

$$\mu = E[X] \qquad \sigma^2 = \mathrm{Var}[X]$$

Standard Normal Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.40 (Standard Normal Distribution)

The *standard normal distribution* is given by $\mu = 0$ and $\sigma^2 = 1$, i.e., $\mathcal{N}(0, 1)$.

Standard Normal Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.40 (Standard Normal Distribution)

The *standard normal distribution* is given by $\mu = 0$ and $\sigma^2 = 1$, i.e., $\mathcal{N}(0, 1)$.

Standard Normal Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Definition 7.40 (Standard Normal Distribution)

The *standard normal distribution* is given by $\mu = 0$ and $\sigma^2 = 1$, i.e., $\mathcal{N}(0, 1)$.

Geometrical Meaning of μ and σ

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

In general:

- μ is the center of the symmetric, bell-shaped curve $f(x)$.
- $\mu + \sigma$ and $\mu - \sigma$ are the inflection points of the curve.

Bivariate Normal Distribution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The joint probability density function of two random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ is

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)}}$$

Multivariate Normal Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

In the general, d-dimensional case for a vector of normal distributed random variables $\vec{X} = (X_1, \ldots, X_d)$, we have the multi-dimensional probability density function for $\vec{x} = (x_1, \ldots, x_d)$

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})^\mathsf{T}}{2}}$$

where $|\Sigma|$ is the determinant of the covariance matrix:

$$\Sigma = \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \mathrm{Cov}(X_2, X_3) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Var}[X_3] & \cdots & \mathrm{Cov}(X_3, X_d) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \mathrm{Cov}(X_d, X_3) & \cdots & \mathrm{Var}[X_d] \end{pmatrix}$$

Standardizing

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- If we have a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable
$$Z = \frac{X - \mu}{\sigma}$$
has a standard normal distribution.
- Standardizing a normal distribution is equivalent to expressing the distance from the mean value as some number of standard deviations (sometimes called the z-score).

Sum of Normal Random Variables

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.41

If $X_1, X_2, \ldots, X_n$ are independent random variables, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$X = \sum_{i=1}^{n} X_i$$

is a random variable with $X \sim \mathcal{N}(\mu, \sigma^2)$ and

$$\mu = E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu_i$$

$$\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = \sum_{i=1}^{n} \sigma_i^2$$

The Central Limit Theorem

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Theorem 7.42 (Central Limit Theorem)

Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables with mean μ and variance $\sigma^2 > 0$. Let Z_n be defined as the standardized average $\overline{X}$:

$$Z_n = \frac{\left(\sum_{i=1}^n X_i\right) - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\left(\sum_{i=1}^n X_i\right) - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Then, in the limit as $n \to \infty$, $Z_n \sim \mathcal{N}(0,1)$ (i.e., $\lim_{n\to\infty} \Pr(Z_n \leq z) = \Phi(z)$, where Φ is the cdf of $\mathcal{N}(0,1)$).

Note that:

The CLT does not assume any particular distribution to be followed by the i.i.d. random variables X_i beyond existing μ and positive variance.

Example: Combination of n Dice

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

adapted from a work by Cmglee [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0) or GFDL (http://www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

Experiment

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Continuous Random Variables
The Uniform Distribution
The Normal Distribution
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

If time allows, let's try this together:

`https://goo.gl/JDaVog`

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
  - Algebraic and Geometric vs. Probabilistic View of Data
  - EM-Clustering
  - Linear Discriminant Analysis
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
  - Algebraic and Geometric vs. Probabilistic View of Data
  - EM-Clustering
  - Linear Discriminant Analysis
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Algebraic and Geometric View of Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- If we have a dataset consisting of d numeric attributes and n data objects, we can describe the data as an $n \times d$ data matrix, where each row is a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d$.
- We can therefore work with the dataset by algebraic or geometric operations such as
  - compute distances and angles between vectors,
  - compute the length of a vector,
  - compute the average of a set of vectors, or
  - transform (multiply, rotate, project) the vectors.
- We follow an algebraic/geometric intuition when working with data in this manner.

Probabilistic View of Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- In the probabilistic view, we interpret each numeric attribute as a random variable.
- Column j in the data matrix is a sequence of n values drawn from the distribution of the corresponding random variable X_j.
- For some j, the observed values x_{ij}, $i = 1, \ldots, n$, can be seen in turn as identity random variables.
- For some j, all x_{ij}, $i = 1, \ldots, n$, can be considered mutually independent and identically distributed as X_j.
- The Central Limit Theorem (Theorem 7.42) therefore supports in many cases the choice of Gaussian distributions to model data adequately.
- However, several classes or clusters in the data means: there are different random variables and/or dependencies between the involved random variables.

Probabilistic View of Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Consequently, we can
  - reason about the distribution followed by each random variable X_j,
  - explore their independence or conditional probabilities,
  - study their joint distributions.
- Overall, we treat the data set at hand as *a sample* drawn from the joint distribution of d random variables.
- A simplifying assumption is the independence of the attributes (cf. Naïve Bayes).
- Note, however, that we can typically not assume that the attribute random variables $X_j, j = 1, \ldots, d$ are identically distributed.

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Compressed overview for an algebraic/geometric view on data: Zaki and Meira Jr. [2014], Chapter 1.3*
- *Compressed overview for a probabilistic view on data: Zaki and Meira Jr. [2014], Chapter 1.4*

Different Interpretations of a Single Algorithm

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We studied k-means clustering and k nearest neighbor classification.
- We introduced both algorithms from an algebraic point of view (computing distances in a vector space).
- Later on, we interpreted k nearest neighbor classification in a probabilistic framework and studied Bayesian learning.
- In this section, we will
  - re-interpret k-means clustering probabilistically and study a more complex clustering approach (EM-clustering), that follows the Bayesian framework;
  - re-interpret Bayesian learning geometrically, leading to the classification method Linear Discriminant Analysis (LDA).

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
  - Algebraic and Geometric vs. Probabilistic View of Data
  - EM-Clustering
  - Linear Discriminant Analysis
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Zaki and Meira Jr. [2014], Chapter 13.*
- *Tan et al. [2006], Chapter 9.2.*

Expectation Maximization

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

EM-clustering [Dempster et al., 1977]:

- objects are points $x = (x_1, \ldots, x_d)$ in Euclidean vector space $\mathbb{R}^d$, $\mathrm{dist} =$ Euclidean distance (L_2)
- cluster is represented by a probability density function
- typically the model is a multivariate normal distribution
- representative for a cluster C:
  - mean vector μ_C of all points in C
  - $d \times d$ covariance matrix Σ_C for the points in C
- probability density of C:

$$f_C(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_C|}} e^{-\frac{1}{2}(x-\mu_C)\cdot(\Sigma_C)^{-1}\cdot(x-\mu_C)^{\mathsf{T}}}$$

Idea of EM clustering

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- each point belongs to several (that is, all) clusters, to each cluster with a different probability $\Pr(x|C)$
- algorithm consists of two alternating steps:
  - assignment of points to their clusters, not absolute but relative, according to probability
  - computation of new cluster representatives
- computation of μ_C needs to consider *relative* assignment
- what is the probability of belonging to some cluster?

Density Function to Model a Cluster

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The model for a cluster C_i is a Gaussian pdf with cluster-specific mean and covariance matrix:

$$f_{C_i}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{C_i}|}} e^{-\frac{1}{2}(x-\mu_{C_i})\cdot(\Sigma_{C_i})^{-1}\cdot(x-\mu_{C_i})^{\mathsf{T}}}$$

Note that:

- *To give probabilities, we would need to compute the integral over a small R around some point x to get an estimate of the probability for some hypothetical point to be located within R, or the relative share of points (e.g., 30%) of the cluster to be in this area.*
- *We therefore do not actually get probabilities here.*
- *However, the values of the p.d.f. are suitable for a relative comparison of probabilities — that is all we actually need here.*

Density Functions to Model Clusters

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- For our purpose, we can use $f_{C_i}(x)$ as a *relative probability*.
- However: this would hold if some point is exclusively assigned to the cluster
- We therefore set the *conditional relative* probability as

$$\Pr(x|C_i) \propto \frac{1}{\sqrt{(2\pi)^d|\Sigma_{C_i}|}} e^{-\frac{1}{2}(x-\mu_{C_i})\cdot(\Sigma_{C_i})^{-1}\cdot(x-\mu_{C_i})^{\mathsf{T}}}$$

Density Functions to Model Clusters

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The overall probability of the data explained by k clusters is:

$$\Pr(x) = \sum_{i=1}^{k} W_i \cdot \Pr(x|C_i)$$

with W_i being the relative share of points of cluster C_i:

$$W_i = \frac{1}{n} \sum_{j=1}^{n} \Pr(C_i|x_j)$$

W_i can be interpreted as probability of the cluster C_i: $\Pr(C_i)$

Conditional Probabilities and Bayes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Bayes' theorem (Theorem 6.8):

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

Hence we have the probability of some point to belong to a given cluster C_i:

$$\Pr(C_i|x) = W_i \cdot \frac{\Pr(x|C_i)}{\Pr(x)}$$

Expectation Maximization: A Clustering Solution as Hypothesis

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- A clustering is a hypothesis to explain the data as generated from a mix of k distributions.
- We search for the best (i.e., the MAP) hypothesis (cf. slide 291).
- We measure the quality of a clustering M as "expectation":

$$E(M) = \sum_{x \in \mathcal{D}} \log(\Pr(x))$$

- The "expectation" characterizes how well the clustering M explains the data.
- The higher $E(M)$, the more likely is it for the dataset to look like it does, given the model M.
- We therefore have to maximize $E(M)$ (cf. slide 303).

Components of the Hypothesis: Gaussian Distributions

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Each of the k multivariate Gaussian distributions is defined by

- mean (vector):

$$\mu_i = \frac{\sum_{x \in \mathcal{D}} x \cdot \Pr(C_i|x)}{\sum_{x \in \mathcal{D}} \Pr(C_i|x)}$$

- covariance matrix of the Gaussian:

$$\Sigma_i = \frac{\sum_{x \in \mathcal{D}} (x - \mu_i) \cdot (x - \mu_i)^\mathsf{T} \cdot \Pr(C_i|x)}{\sum_{x \in \mathcal{D}} \Pr(C_i|x)}$$

Expectation Maximization Clustering Algorithm

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.1 (EM-Clustering [Dempster et al., 1977])

Given: $\mathcal{D}$, $k \in \mathbb{N}$
create initial model $M = (C_1, \ldots, C_k)$;
for *all* $x \in \mathcal{D}$ *and all* $C_i \in M$:
  compute $\Pr(x|C_i)$, $\Pr(x)$, $\Pr(C_i|x)$;
$EM := E(M)$;
repeat
  compute new model $M = (C_1, \ldots, C_k)$
    with the previous W_i, μ_i, *and* Σ_i;
  for *all* $x \in \mathcal{D}$ *and all* $C_i \in M$:
    compute $\Pr(x|C_i)$, $\Pr(x)$, $\Pr(C_i|x)$;
  $EM' := E(M)$;
  $\delta := |EM' - EM|$;
  $EM := EM'$;
until $\delta < \varepsilon$;
return M;

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 0):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 1):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 2):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 3):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 4):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 5):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 6):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 7):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 8):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 9):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 10):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 11):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 12):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 13):

EM Example: Iterative Fitting of the Model Parameters μ_i and Σ_i

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Old Faithful geyser data (Iteration 14):

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- $\mathcal{O}(n \cdot k \cdot \#\text{iterations})$ – usually many iterations required
- Result quality and runtime depend strongly on initial model and good choice of k.
- Modification for returning k disjunct clusters: each object belongs only to the cluster with highest (relative) probability $\Pr(x|C_i)$.

Multivariate Normal Distribution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Clusters can be:

- Spherical (A)
- Ellipsoid (C)
- Rotated ellipsoid (B)

But: same formula for each situation!

Probability density function of multivariate normal distribution:

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^{\mathsf{T}}\right)}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\mathsf{T}\right)}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\mathsf{T}\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\top\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\top\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

normalization and squared distance from mean

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\top\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

normalization and squared distance from mean

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\top\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x,\mu,\Sigma) := \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)^\top}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\mathsf{T}\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x,\mu,\Sigma)^2 := (x-\mu)\Sigma^{-1}(x-\mu)^\mathsf{T}$$

Multivariate Normal Distribution

Dissecting the formula

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$f(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)\Sigma^{-1}(p-\mu)^\mathsf{T}\right)}$$

Uni-variate normal distribution

$$f(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x,\mu,\Sigma)^2 := (x-\mu)\Sigma^{-1}(x-\mu)^\mathsf{T}$$

(compare the general quadratic form distance, Slide 119)

What is the effect of Σ^{-1}?

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The covariance matrix is symmetric, and non-negative on the diagonal. Therefore it can usually be inverted (ignore constant dimensions).

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The covariance matrix is symmetric, and non-negative on the diagonal. Therefore it can usually be inverted (ignore constant dimensions).
Decomposition (equivalently):

$$V\Lambda V^{-1} = \Sigma \quad \equiv \quad V\Lambda^{-1}V^{-1} = \Sigma^{-1}$$

- V contains the eigenvectors
- Λ contains the eigenvalues on the diagonal.

Note that:

$V \cong$ rotation, $\Lambda \cong$ scaling!
This is the key idea of Principal Component Analysis.

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)\Sigma^{-1}(x-\mu)^{\mathsf{T}}$$

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)V\Omega(V\Omega)^{\mathsf{T}}(x-\mu)^{\mathsf{T}}$$

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)V\Omega(V\Omega)^{\mathsf{T}}(x-\mu)^{\mathsf{T}} \\ &= \langle (V\Omega)^{\mathsf{T}}(x-\mu), (V\Omega)^{\mathsf{T}}(x-\mu)\rangle \end{aligned}$$

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)V\Omega(V\Omega)^{\mathsf{T}}(x-\mu)^{\mathsf{T}} \\ &= \langle (V\Omega)^{\mathsf{T}}(x-\mu), (V\Omega)^{\mathsf{T}}(x-\mu)\rangle \\ &= L_2((V\Omega)^{\mathsf{T}}(x-\mu))^2 \end{aligned}$$

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^\mathsf{T}V^\mathsf{T} = V\Omega(V\Omega)^\mathsf{T}$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)V\Omega(V\Omega)^\mathsf{T}(x-\mu)^\mathsf{T} \\ &= \langle (V\Omega)^\mathsf{T}(x-\mu), (V\Omega)^\mathsf{T}(x-\mu)\rangle \\ &= L_2((V\Omega)^\mathsf{T}(x-\mu))^2 \end{aligned}$$

norm $L_2(x-y) = \mathrm{dist}_{L_2}(x,y)$ Euclidean distance

Multivariate normal distribution

Inverse of covariance matrix – Σ^{-1}

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Build Ω using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^{\mathsf{T}}V^{\mathsf{T}} = V\Omega(V\Omega)^{\mathsf{T}}$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)V\Omega(V\Omega)^{\mathsf{T}}(x-\mu)^{\mathsf{T}} \\ &= \langle (V\Omega)^{\mathsf{T}}(x-\mu), (V\Omega)^{\mathsf{T}}(x-\mu)\rangle \\ &= L_2((V\Omega)^{\mathsf{T}}(x-\mu))^2 \end{aligned}$$

norm $L_2(x-y) = \mathrm{dist}_{L_2}(x,y)$ Euclidean distance

Note that:

The Mahalanobis distance can be interpreted as Euclidean distance after having applied PCA on the data.

Model

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- (relative) assignment of points to closest cluster representative
- "closest" is given by Mahalanobis distance:
  - quadratic form distance
  - distance matrix depends on the corresponding cluster (covariance matrix of the cluster points)

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Expectation Maximization

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

From the General Principle to the Specialized Solution

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
  - Algebraic and Geometric vs. Probabilistic View of Data
  - EM-Clustering
  - Linear Discriminant Analysis
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Multivariate Gaussian Model for Each Class

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Classification problem, given training data $TR \subseteq \mathbb{R}^d$:

- model each class $c_i \in C$ as a multivariate Gaussian:

$$f_{c_i}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{c_i}|}} e^{-\frac{1}{2}(x-\mu_{c_i})\cdot(\Sigma_{c_i})^{-1}\cdot(x-\mu_{c_i})^\mathsf{T}}$$

- estimate the model parameters mean μ_{c_i} and covariance matrix (Σ_{c_i}) for each class c_i from the training data
- let $C_i \subseteq TR$ be the set of vectors belonging to class c_i:

$$\mu_{c_i} = \frac{\sum_{x \in C_i} x}{|C_i|}$$

$$(\Sigma_{c_i})_{mn} = \frac{\sum_{x \in C_i} (x_m - \mu_{c_i m}) \cdot (x_n - \mu_{c_i n})}{|C_i|}$$

Maximum Likelihood Prediction

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Given the Gaussian model for each class, we can predict the maximum likelihood class (cf. slide 291):

$$
\begin{aligned}
h(x) &= \arg\max_{c_i \in C} \Pr(c_i|x) \\
&= \arg\max_{c_i \in C} \Pr(x|c_i) \cdot \Pr(c_i) \\
&= \arg\max_{c_i \in C} \frac{1}{\sqrt{(2\pi)^d |\Sigma_{c_i}|}} e^{-\frac{1}{2}(x-\mu_{c_i})\cdot(\Sigma_{c_i})^{-1}\cdot(x-\mu_{c_i})^{\mathsf{T}}} \cdot \Pr(c_i)
\end{aligned}
$$

Global Covariance Matrix

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Problem: sometimes the amount of available training examples per class is not sufficient to significantly determine the covariance matrix for each class.
- Combine the covariance matrices to a global covariance matrix:

$$(\Sigma) = \frac{\sum_{c_i \in C} (\Sigma_{c_i})}{|C|}$$

- classification rule:

$$h(x) = \arg\max_{c_i \in C} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu_{c_i})\cdot(\Sigma)^{-1}\cdot(x-\mu_{c_i})^{\mathsf{T}}} \cdot \Pr(c_i)$$

Linear Discriminant Analysis

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Data: Algebraic vs. Probabilistic
EM-Clustering
LDA
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$
\begin{aligned}
h(x) &= \arg\max_{c_i \in C} \left\{ \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu_{c_i})\cdot(\Sigma)^{-1}\cdot(x-\mu_{c_i})^\mathsf{T}} \cdot \Pr(c_i) \right\} \\
&= \arg\max_{c_i \in C} \left\{ -\frac{1}{2}(x-\mu_{c_i})\cdot(\Sigma)^{-1}\cdot(x-\mu_{c_i})^\mathsf{T} + \log \Pr(c_i) \right\} \\
&= \arg\max_{c_i \in C} \left\{ x(\Sigma)^{-1}\mu_{c_i}^\mathsf{T} - \frac{1}{2}\mu_{c_i}\cdot(\Sigma)^{-1}\mu_{c_i}^\mathsf{T} + \log \Pr(c_i) \right\} \\
&= \arg\max_{c_i \in C} \sigma_{c_i}(x)
\end{aligned}
$$

- Compute only the linear discriminant function $\sigma_{c_i}(x)$ for each c_i, no actual probabilities.
- As only the means are different, we have a linear separation.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
  - Non-Parametric Density Estimation
  - Density-based Clustering
  - Hierarchical Clustering
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Parametric vs. Non-Parametric Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

parametric learning Given a data sample, we assume it follows a specific family of distributions (e.g., Gaussian).

- We fit a Gaussian distribution to the data sample.
- "'Fitting" a distribution means, we learn the parameters of the distribution (in the example of a Gaussian, we learn μ and Σ).

non-parametric learning Given a data sample, we try to infer probability estimates directly from the sample.

- We might approximate the shape of a density distribution without knowing its family.
- We do not necessarily need to derive an *explicit* density function in order to estimate probabilities.

Examples: Parametric Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- k-means (estimate μ_i of k Gaussians with unit variance and zero covariance)
- EM-clustering (estimate μ_i and Σ_i of k Gaussians)

- LDA (estimate μ_i and Σ of k Gaussians and of the complete data set)

Examples: Non-Parametric Learning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- k nearest neighbor classifier (approximate class distribution boundaries, i.e., the shapes of implicit density distribution functions)

- naïve Bayes classifier (approximate probability estimates without deriving an explicit density function)

In both cases, we assume there is an underlying probability density distribution function f. We approximate this function f only locally but do not conclude on its shape overall.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
  - Non-Parametric Density Estimation
  - Density-based Clustering
  - Hierarchical Clustering
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

The Problem of Density Estimation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The point-wise, local approximation of a probability density distribution function is more generally treated as non-parametric "density estimation".

Recommended Reading:

- *The classic text on non-parametric density estimation is by Silverman [1986].*
- *Deep technical coverage by Duda et al. [2001], Chapter 4.*
- *Concise and illustrative introduction by Zaki and Meira Jr. [2014], Chapter 15.2. (Their material is used here.)*

Problem Statement

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- We are given a data set consisting of the n points $x_1, \ldots, x_n \in \mathbb{R}^d$.
- We assume that these points are a sample of values assumed by some random variable $X \sim f(x)$ (probabilistic point of view).
- We do not know the underlying probability density distribution function f in general.
- We would like to know the value $f(x)$ for some given x.

Estimation of the Cumulative Distribution Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- In a one-dimensional scenario (i.e., $x_i \in \mathbb{R}$), the cumulative density distribution function can be directly estimated from the data.
- We have to count the points that are less than or equal to x:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

with $I(b) = \begin{cases} 1 & \text{if } b \\ 0 & \text{if } \neg b \end{cases}$

Estimation of the Probability Density Function Around x from the CDF

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- From the estimate $\hat{F}$ of F, we can estimate the density function f by the derivative of $\hat{F}$.
- We take a small window of width h, centered at x:

$$\hat{f}(x) = \frac{\hat{F}\left(x + \frac{h}{2}\right) - \hat{F}\left(x - \frac{h}{2}\right)}{h} = \frac{k/n}{h} = \frac{k}{nh} \tag{7.2}$$

  where k is the number of points located within the window of size h around x.
- The density estimate is therefore given as the ratio of the fraction of points within the window $\frac{k}{n}$ to the volume h of the window.
- Choice of h:
  - Too large will consider many neighbor points and smoothen the estimated function.
  - Too small will not take enough points and does not give an accurate estimate.

Kernel Estimator

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Kernel density estimation uses a *kernel function* K.
- A kernel function is
  - non-negative, i.e., $\forall x : K(x) \geq 0$,
  - symmetric, i.e., $\forall x : K(-x) = K(x)$, and
  - integrates to 1, i.e., $\int K(x)\, dx = 1$.
- Therefore K qualifies as a probability density function itself.

Discrete Kernel

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The discrete kernel simply counts the number of points in a window of width h:

$$K(z) = \begin{cases} 1 & \text{if } |z| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

- If $|z| = \left|\frac{x-x_i}{h}\right| \leq \frac{1}{2}$, the point x_i is within the window of width h centered at x.
- We can rewrite the density estimate from Eq. 7.2 by using the discrete kernel:

$$\begin{aligned} \hat{f}(x) &= \frac{k}{nh} \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \end{aligned}$$

Kernel Density Estimation with Discrete Kernel

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Zaki and Meira Jr. [2014].

Gaussian Kernel

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- In the discrete kernel, we have an all-or-nothing contribution of the points around x to the estimate:
  - If some point falls in the window around x, it contributes $\frac{1}{hn}$ to the probability estimate $\hat{f}(x)$.
  - If some point is outside the window, it contributes 0.
- A Gaussian kernel results in a smoother change of the influence of some point:

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

- The center of the window, x, relates to the mean, window size h relates to the standard deviation.

KDE with Gaussian Kernel

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Zaki and Meira Jr. [2014].

Multivariate Density Estimation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- To estimate the density around a d-dimensional point $x \in \mathbb{R}^d$, we replace the one-dimensional window by a d-dimensional hypercube.
- Instead of the window length h, we need to account for the volume h^d.
- With some multivariate kernel function K, the density estimate becomes:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Multivariate Kernel Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- The discrete kernel function generalizes directly to a d-dimensional vector space:

$$K(z) = \begin{cases} 1 & \text{if } |z_j| \leq \frac{1}{2} \text{ for all dimensions } j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

- For the d-dimensional Gaussian kernel, we assume that the covariance matrix is the identity matrix:

$$K(z) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{zz^{\mathsf{T}}}{2}}$$

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{(x-x_i)(x-x_i)^{\mathsf{T}}}{2h^2}}$$

Two-dimensional Gaussian KDE

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Zaki and Meira Jr. [2014].

Example Data – Gaussian KDE, $h = 20$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Zaki and Meira Jr. [2014].

Example Data – Gaussian KDE, $h = 20$

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Zaki and Meira Jr. [2014].

Nearest Neighbor Density Estimation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- So far we kept the volume fixed by a fixed window width (or hypercube edge length) h.
- Problem: how to choose h in order to capture enough, but not too many points?
- Alternative: fix the number k of points, allow the volume to vary in order to capture k points.
- With Euclidean distance, we get the volume $V_k(x)$ of a d-dimensional hypersphere with the distance from x to its kth nearest neighbor as radius.
- The resulting k nearest neighbor density estimation is:

$$\hat{f}(x) = \frac{k}{nV_k(x)}$$

- Cf. the probabilistic reasoning on the $k$nn classifier (slide 285)!

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
  - Non-Parametric Density Estimation
  - Density-based Clustering
  - Hierarchical Clustering
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *The original paper on DBSCAN [Ester et al., 1996].*
- *Survey on density-based clustering [Kriegel et al., 2011].*
- *Tan et al. [2006], Chapters 8.4, 9.3, 9.4.5–9.4.8.*

General Idea

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Clusters are high-density “areas” (volumes) in a d dimensional space.
- Clusters are separated by low-density “areas” (volumes).
- The high-density area defined by the set of objects forming a cluster is connected in the data space.

Central Assumption for Density-based Clusters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Assumption 7.1 (Density-based Cluster)

For each object in a cluster, the local density exceeds a given threshold:

$$\forall o \in C : \phi(o) \geq \theta$$

Cf. density-contour by Hartigan [1975] and discussion by Kriegel et al. [2011], Campello et al. [2015].

Non-Parametric Clustering: Intuition [Kriegel et al., 2011]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

parametric learning
approximate some density function of a given family (e.g. Gaussian)

non-parametric no assumption on family of density function, identify areas of high density

Intuition: High Density Threshold

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- high density threshold: cluster of lower density is regarded as noise

Intuition: Medium Density Threshold

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- medium density threshold: three clusters are identified, some points in the tails are regarded as noise

Intuition: Low Density Threshold

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- low density threshold: two clusters are merged

Partitioning vs. Density-based Clustering

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Partitioning approaches solve an optimization problem:
  - many possible solutions of different quality
  - find the (local!) optimum
  - no guarantee or knowledge about the truly best solution
- The true density-based cluster solution is uniquely defined (in theory).
  - Solutions differ due to
    - algorithmic strategies
    - definitions of neighborhood
    - density estimation techniques

Intuition: “Natural” Clusters

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Density-based clusters make often sense in natural structures, such as geographic, biological data.
- Model simplicity: we could express any data sample as a combination of Gaussians, but not assuming a Gaussian (or any particular distribution family) might often yield simpler, more elegant solutions.
  - Recall the sparsity principle, slide 299.
- Example: topological profile of Mt. Eden (a volcano in Auckland, NZ)

Intuition: “Natural” Clusters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Example: different species of iris
- Within a species, we have a certain variation but no jumps

Intuition: “Natural” Clusters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Intuition: “Natural” Clusters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Shell coiling in three dimensions [Raup, 1966]:

TEXT-FIG. 2—Computer produced shell forms illustrating the morphological effects of change in whorl expansion rate (W), translation rate (T), and the distance from the generating curve to the axis (D),

Intuition: “Natural” Clusters

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Shell coiling in three dimensions [Raup, 1966]:

TEXT-FIG. 4—Three dimensional block illustrating the spectrum of possible shell forms. The shape of the generating curve is assumed to be constant. The regions occupied by the majority of species in four taxonomic groups are outlined in the block. Species of these groups are not commonly found in the blank regions of the block.

Density-based Clustering: Basic Concepts

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A density threshold is defined by parameters ε and `MinPts`

Density-based Clustering: Basic Concepts

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A density threshold is defined by parameters ε and `MinPts`

- core point

Density-based Clustering: Basic Concepts

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A density threshold is defined by parameters ε and `MinPts`

- core point
- direct density-reachability

Density-based Clustering: Basic Concepts

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A density threshold is defined by parameters ε and `MinPts`

- core point
- direct density-reachability
- density-reachability

Density-based Clustering: Basic Concepts

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A density threshold is defined by parameters ε and `MinPts`

- core point
- direct density-reachability
- density-reachability
- density-connectivity

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\mathrm{RQ}(p, \varepsilon) = \{o \in \mathcal{D} \mid \mathrm{dist}(p, o) \leq \varepsilon\}$$

- $p \in \mathcal{D}$ is a *core point* if: $|\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\mathrm{RQ}(p, \varepsilon) = \{o \in \mathcal{D} \mid \mathrm{dist}(p, o) \leq \varepsilon\}$$

- $p \in \mathcal{D}$ is a *core point* if: $|\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$
- $q \in \mathcal{D}$ is directly density-reachable from $p \in \mathcal{D}$ w.r.t. ε and MinPts if: $q \in \mathrm{RQ}(p, \varepsilon) \wedge |\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\mathrm{RQ}(p, \varepsilon) = \{o \in \mathcal{D} \mid \mathrm{dist}(p, o) \leq \varepsilon\}$$

- $p \in \mathcal{D}$ is a *core point* if: $|\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$
- $q \in \mathcal{D}$ is directly density-reachable from $p \in \mathcal{D}$ w.r.t. ε and MinPts if: $q \in \mathrm{RQ}(p, \varepsilon) \wedge |\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$
- $q \in \mathcal{D}$ is density-reachable from $p \in \mathcal{D}$ w.r.t. ε and MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = p$, $p_n = q$, and p_{i+1} is directly density-reachable from p_i

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$\mathrm{RQ}(p, \varepsilon) = \{o \in \mathcal{D} \mid \mathrm{dist}(p, o) \leq \varepsilon\}$$

- $p \in \mathcal{D}$ is a *core point* if: $|\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$
- $q \in \mathcal{D}$ is directly density-reachable from $p \in \mathcal{D}$ w.r.t. ε and MinPts if: $q \in \mathrm{RQ}(p, \varepsilon) \wedge |\mathrm{RQ}(p, \varepsilon)| \geq \mathrm{MinPts}$
- $q \in \mathcal{D}$ is density-reachable from $p \in \mathcal{D}$ w.r.t. ε and MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = p$, $p_n = q$, and p_{i+1} is directly density-reachable from p_i
- $p \in \mathcal{D}$ is density-connected to $q \in \mathcal{D}$ w.r.t. ε and MinPts if $\exists o \in \mathcal{D}$ such that both, p and q are density-reachable from o w.r.t. ε and MinPts.

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A *(density-based) cluster* C w.r.t. ε and MinPts is a non-empty subset of $\mathcal{D}$ satisfying the following conditions:

1. *maximality*: $\forall p, q$: if $p \in C$ and q is density-reachable from p w.r.t. ε and MinPts, then $q \in C$.
2. *connectivity*: $\forall p, q \in C$: p is density-connected to q w.r.t. ε and MinPts.

Formalization [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A *(density-based) cluster* C w.r.t. ε and MinPts is a non-empty subset of $\mathcal{D}$ satisfying the following conditions:

1. *maximality*: $\forall p, q$: if $p \in C$ and q is density-reachable from p w.r.t. ε and MinPts, then $q \in C$.
2. *connectivity*: $\forall p, q \in C$: p is density-connected to q w.r.t. ε and MinPts.

Note that:

This fundamental property enables efficient search for clusters!

DBSCAN [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.2 (DBSCAN [Ester et al., 1996])

```
DBSCAN(DB,Eps,MinPts)
  // DB is UNCLASSIFIED
  C_Id := nextId(NOISE);
  FOR i FROM 1 TO DB.size DO
    Point := DB.get(i);
    IF Point.C_Id = UNCLASSIFIED THEN
      IF ExpandCluster(DB, Point, C_Id, Eps, MinPts) THEN
        C_Id := nextId(C_Id)
```

DBSCAN [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.3 (DBSCAN – expandCluster)

```
ExpandCluster(DB, Point, C_Id, Eps, MinPts):  Boolean
  seeds := DB.rq(Point,Eps)
  IF seeds.size < MinPts THEN // no core point
    DB.changeC_Id(Point,NOISE);
    RETURN FALSE;
  ELSE // all points in seeds are dens-reach from Point
    DB.changeC_Ids(seeds,C_Id);
    seeds.delete(Point);
  WHILE seeds <> Empty DO
    currentP := seeds.first();
    result := DB.rq(currentP,Eps);
    IF result.size >= MinPts THEN
      FOR i FROM 1 TO result.size DO
        resultP := result.get(i);
        IF resultP.C_Id IN {UNCLASSIFIED,NOISE} THEN
          IF resultP.CiId = UNCLASSIFIED THEN
            seeds.append(resultP);
          DB.changeC_Id(resultP,C_Id);
    seeds.delete(currentP);

  RETURN TRUE;
```

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
BC

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
B

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
BD

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
D

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
D

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
G

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
H

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
IJ

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
I

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
I

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
O

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPR

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NP

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPQST

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPQS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPQS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
NPS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
PS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
PS

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
S

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:
S

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

Example 1

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

No border points!

Example 1

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 2$

Seed list:

No border points!

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
BC

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
B

Example 2

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
BD

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
D

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
D

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
H

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
IJ

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
I

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
I

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
NPR

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
NP

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list: NPQST

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
NPQS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list: NPQS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
NPS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
NPS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
PS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
PS

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
S

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:
S

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 2

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 3$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
IJ

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
I

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
I

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPR

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NP

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPQST

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPQS

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPQS

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPS

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
NPS

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
PS

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
PS

Example 3

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
S

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:
S

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 3

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 1.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
BCD

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
BD

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
BD

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
D

Example 4

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
D

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
HIJ

Example 4

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
IJ

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
IJ

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
I

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
IK

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
K

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
K

Example 4

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list: ONPR

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
NPR

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list: NPRQST

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
NPQST

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
NPQST

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list: NPQS

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list: NPQS

Example 4

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
NPS

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
NPS

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
PS

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
PS

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
S

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:
S

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 4

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 2.1$

$minPts = 4$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
E

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list: GHIJ

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
HIJ

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
HIJK

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
IJK

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
IJK

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
IK

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
IK

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
K

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
K

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
L

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list: ONPRQST

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPQST

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPQST

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPQST

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPQST

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPQS

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list: NPQS

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
NPS

Example 5

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list: NPS

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
PS

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
PS

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
S

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:
S

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

Example 5

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 5$

Seed list:

DBSCAN [Ester et al., 1996]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.3 (DBSCAN – expandCluster)

```
ExpandCluster(DB, Point, C_Id, Eps, MinPts):  Boolean
  seeds := DB.rq(Point,Eps)
  IF seeds.size < MinPts THEN // no core point
    DB.changeC_Id(Point,NOISE);
    RETURN FALSE;
  ELSE // all points in seeds are dens-reach from Point
    DB.changeC_Ids(seeds,C_Id);
    seeds.delete(Point);
  WHILE seeds <> Empty DO
    currentP := seeds.first();
    result := DB.rq(currentP,Eps);
    IF result.size >= MinPts THEN
      FOR i FROM 1 TO result.size DO
        resultP := result.get(i);
        IF resultP.C_Id IN {UNCLASSIFIED,NOISE} THEN
          IF resultP.CiId = UNCLASSIFIED THEN
            seeds.append(resultP);
          DB.changeC_Id(resultP,C_Id);
    seeds.delete(currentP);

  RETURN TRUE;
```

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:
D

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:
E

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:
K

Example 6

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:
L

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:
MO

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

Example 6

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$\varepsilon = 4.1$

$minPts = 4$

Seed list:

single-link like effect!

DBSCAN* [Campello et al., 2013a, 2015]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

These papers tackle hierarchical clustering – but provide as a prerequisite a redefinition of DBSCAN, excluding border points from any cluster.

Note that:

Only then the central assumption of density-based clustering (Assumption 7.1) holds:

> *For each object in a cluster, the local density exceeds a given threshold.*

Shared Nearest Neighbor (SNN) Clustering

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- DBSCAN can detect clusters of different size and shape
- but cannot find clusters for different density thresholds.

Improvement: different notion of similarity

- similarity of objects, if both are close to a reference set
- similarity is “approved” by reference set
- example: reference set is set of common neighbors, similarity is cardinality of this set
- shared nearest neighbor (SNN) similarity ($\mathrm{NN_k}(o)$: set of the k nearest neighbors of o):

$$\mathrm{SNN}_k(p,q) = |\,\mathrm{NN_k}(p) \cap \mathrm{NN_k}(q)|$$

SNN_6-similarity(●,●) = 4

SNN_6-similarity(●,●) = 0

See also studies by Houle et al. [2010], Bernecker et al. [2011] on the quality improvement in similarity search by using SNN.

SNN clustering [Jarvis and Patrick, 1973]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.4 (SNN clustering [Jarvis and Patrick, 1973])

1. *compute similarity matrix and similarity graph*
   - *for all pairs $o, p \in \mathcal{D}$: compute SNN_k similarity*
   - *SNN similarity graph: nodes=objects, edges weight=SNN similarity*
   - *no edge where SNN similarity = 0*
2. *generate clusters*
   - *delete edges with weight below threshold τ*
   - *cluster: connected component in the resulting graph*

database ⇒ SNN similarity graph ⇒ generate clusters

Problems of the Algorithm of Jarvis and Patrick [1973]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Problem:

- hard to find a good threshold τ
- small variations lead to very different solutions

picture from Ertöz et al. [2003]

Solution of Ertöz et al. [2003]:

- combine SNN similarity with density-based concepts
- SNN density: number of points within a given radius ε w.r.t. SNN similarity

$$\mathrm{SNN}_k\text{-density}(p, \varepsilon) = |\{q | \mathrm{SNN}_k(p, q) \geq \varepsilon\}|$$

SNN clustering [Ertöz et al., 2003]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(a) All Points (b) High SNN Density

(c) Medium SNN Density (d) Low SNN Density

picture from Ertöz et al. [2003]

SNN clustering [Ertöz et al., 2003]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.5 (SNN clustering [Ertöz et al., 2003])

given k, ε, MinPts*:*

1. *compute similarity matrix and -graph (as for Alg. 7.4)*
2. *compute SNN density for each point w.r.t.* ε
3. *derive core points w.r.t.* MinPts
4. *cluster core points* p, q *if* $\mathrm{SNN}_k(p,q) \geq \varepsilon$
5. *assign a non-core point* p *to cluster* C *if there is a core point* $q \in C$ *with* $\mathrm{SNN}_k(p,q) \geq \varepsilon$
6. *all remaining non-core points are noise*

SNN clustering [Ertöz et al., 2003]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.5 (SNN clustering [Ertöz et al., 2003])

given k, ε, MinPts*:*

1. *compute similarity matrix and -graph (as for Alg. 7.4)*
2. *compute SNN density for each point w.r.t.* ε
3. *derive core points w.r.t.* MinPts
4. *cluster core points* p, q *if* $\mathrm{SNN}_k(p,q) \geq \varepsilon$
5. *assign a non-core point* p *to cluster* C *if there is a core point* $q \in C$ *with* $\mathrm{SNN}_k(p,q) \geq \varepsilon$
6. *all remaining non-core points are noise*

Note that:

Steps 2 to 6 are nothing but DBSCAN when using a different notion of distance!

Generalization of DBSCAN [Sander et al., 1998]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

basic idea of DBSCAN:
ε-neighborhood contains at least MinPts points

$\swarrow$ $\qquad$ $\searrow$

$\mathrm{dist} \leq \varepsilon$ $\qquad$ $|\,\mathrm{RQ}(o, \varepsilon)| \geq \mathrm{MinPts}$

$\downarrow$ $\qquad$ $\downarrow$

$\mathrm{NPred}(o, p)$ $\qquad$ $\mathrm{MinWeight}(N)$

reflexive, symmetric for pairs of objects $\qquad$ arbitrary predicate for set of objects

$\downarrow$ $\qquad$ $\downarrow$

generalized neighborhood $\qquad$ generalized minimal cardinality

$N_{\mathrm{NPred}}(o) = \{p|\,\mathrm{NPred}(o, p)\}$ $\qquad$ $\mathrm{MinWeight}(N_{\mathrm{NPred}}(o))$

$\searrow$ $\qquad$ $\swarrow$

NPred neighborhood has at least “weight” MinWeight.

Examples

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

NPred	$\mathrm{dist}(p,q) \leq \varepsilon$	intersect(p,q)	neighbor cell and similar color
MinWeight	cardinality(...) $\geq$ MinPts	sum of area $\geq$ 5% of total area	`true`

Algorithm GDBSCAN [Sander et al., 1998]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- same algorithmic schema as DBSCAN
- instead of $\mathrm{RQ}(o, \varepsilon)$: N_{NPred}-query
- instead of condition $|\mathrm{RQ}(o, \varepsilon)| \geq \mathrm{MinPts}$: evaluate $\mathrm{MinWeight}$ predicate
- almost arbitrary NPred possible (reflexive and symmetric)

Note that:

Complexity depends mainly on the complexity of the N_{NPred}-query.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
  - Non-Parametric Density Estimation
  - Density-based Clustering
  - Hierarchical Clustering
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Recommended Reading

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *Tan et al. [2006], Chapter 8.3.*
- *Zaki and Meira Jr. [2014], Chapter 14.*
- *Discussion of the field by Campello et al. [2015].*

General Idea

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Instead of deriving a partition of the dataset into a flat set of clusters, derive a hierarchy (dendrogram) of clusters.
- A cluster at a higher level can contain several smaller clusters at a lower level.
- A tree represents a hierarchical clustering:
  - root: represents the complete database
  - leaf: represents a single object
  - inner node: represents a cluster containing all the objects of the subtree rooted at this node

Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

categories of hierarchical clustering algorithms (typically best-first heuristic):

- bottom-up construction of the dendrogram (agglomerative), at each node merging two clusters
- top-down construction of the dendrogram (divisive) at each node splitting a cluster into two parts

Agglomerative Algorithm

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Algorithm 7.6 (Agglomerative Hierarchical Clustering)

1. *Initial clusters consist of one object each.*
2. *Compute distances between all pairs of clusters.*
3. *Merge those two clusters with smallest distance.*
4. *Compute distance between the new cluster and all remaining clusters.*
5. *If only one cluster remains: terminate; otherwise: repeat from step 3.*

Complexity

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

initialization:

similarity matrix $\mathcal{O}(n^2)$

(pairwise similarity between n objects)

iteration i:

build a new cluster $\mathcal{O}((n-i+1)^2)$

compute new similarities $\mathcal{O}(n-i+1)$

n iterations $\Rightarrow \mathcal{O}(n^3)$

Distances Between Clusters

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Different approaches are based on different distance measures.
- Using a particular distance measure can allow for algorithmic enhancements.
- Some common distance measures are:
  - single link
  - complete link
  - average link (a.k.a.: UPGMA – unweighted pair group method with arithmetic mean)

Single Link

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Let $\mathrm{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_0^+$ be a distance measure for pairs of objects, let $X, Y \subseteq \mathcal{D}$ designate sets of objects.

$$\mathrm{dist}_{\text{single link}}(X, Y) = \min_{x \in X, y \in Y} \mathrm{dist}(x, y)$$

- efficient algorithm: SLINK [Sibson, 1973]: $\mathcal{O}(n^2)$
- single-link-effect: chain-like clusters, due to the merging of clusters through some chain of objects, yields:
  - clusters with high variance
  - clusters with an extended structure

Complete Link

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Let $\mathrm{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_0^+$ be a distance measure for pairs of objects, let $X, Y \subseteq \mathcal{D}$ designate sets of objects.

$$\mathrm{dist}_{\text{complete link}}(X, Y) = \max_{x \in X, y \in Y} \mathrm{dist}(x, y)$$

- efficient algorithm: CLINK [Defays, 1977]: $\mathcal{O}(n^2)$
- complete-link-effect:
  - small, strongly delimited clusters
  - convex clusters with similar diameters

Average Link

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Let $\mathrm{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_0^+$ be a distance measure for pairs of objects, let $X, Y \subseteq \mathcal{D}$ designate sets of objects.

$$\mathrm{dist}_{\text{average link}}(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} \mathrm{dist}(x, y)$$

Average-Link

- sort of a compromise between single link and complete link
- no efficient algorithm known, i.e., $\mathcal{O}(n^3)$

Average Link

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Let $\mathrm{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_0^+$ be a distance measure for pairs of objects, let $X, Y \subseteq \mathcal{D}$ designate sets of objects.

$$\mathrm{dist}_{\text{average link}}(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} \mathrm{dist}(x, y)$$

Average-Link

- sort of a compromise between single link and complete link
- no efficient algorithm known, i.e., $\mathcal{O}(n^3)$

Note that:

Fundamental properties of distance measures do not hold for some of these distance measures – think about it!

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example: Single Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A B C D E F G H J I K L M O N Q P S R T

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example | Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Complete Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

A B C D E F G H J I K L M O N P Q S R T

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example Average Link Clustering

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

pros

- Does not require knowledge of number of clusters.
- Not only flat partition but a hierarchy of clusters.
- A single partition can be retrieved by some horizontal cut.

cons

- If you want to have a flat partition, where is a good place to cut?
- Greedy heuristic, cannot correct bad decisions.
- single-link-effect, complete-link-effect
- in general: inefficient

Density-based Hierarchical Clustering

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- observation: for a constant value of MinPts, density-based clusters with a smaller value of ε are completely contained within clusters with a larger value of ε

- idea: in a DBSCAN-like run compute simultaneously clustering for different density-thresholds:
  - first high density clusters
  - then clusters with lower density
- design a representation different from a dendrogram, to remain readable for large datasets
- cluster model: density-based cluster

OPTICS [Ankerst et al., 1999]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Basic concepts:

- core distance of an object o w.r.t. $\varepsilon, \mathrm{MinPts}$:

$$\mathrm{coredist}_{\varepsilon,\mathrm{MinPts}}(o) = \begin{cases} \text{undefined, if } |\mathrm{RQ}(o,\varepsilon)| < \mathrm{MinPts} \\ \mathrm{dist}_{\mathrm{MinPts}}(o), \text{ else} \end{cases}$$

- reachability distance of object p w.r.t. object o:

$$\mathrm{reachdist}_{\varepsilon,\mathrm{MinPts}}(p,o) = \begin{cases} \text{undefined, if } |\mathrm{RQ}(o,\varepsilon)| < \mathrm{MinPts} \\ \max\{\mathrm{coredist}(o), \mathrm{dist}(o,p)\}, \text{ else} \end{cases}$$

OPTICS [Ankerst et al., 1999]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Result: cluster order

- OPTICS does not directly deliver a (hierarchical) clustering but a "cluster order" w.r.t. ε and MinPts
- starts with an arbitrary object
- continue with the object exhibiting minimal $\mathrm{reachdist}$ to any of the previously seen objects

OPTICS [Ankerst et al., 1999]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

data structures:

- SeedList (keeps visited points sorted w.r.t. current reachability distance)
- ClusterOrder (result, is built successively)

Algorithm 7.7 (OPTICS [Ankerst et al., 1999])

```
SeedList := ∅;
WHILE there are unlabeled objects in 𝒟 DO
  IF SeedList = ∅
  THEN
    insert arbitrary object into SeedList with reachdist = ∞;
  ELSE
    insert first object ("o") from SeedList
      with current reachdist into ClusterOrder;
    label o as processed;
  FOR ALL neighbor ∈ RQ(o, ε) DO
    SeedList.update(neighbor,o);
```

OPTICS [Ankerst et al., 1999]

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

insert/update object in SeedList:

- for each object p in SeedList we keep the current $\mathrm{reachdist}$: p.rdist
- SeedList is organized as a heap (sorted with increasing rdist)

Algorithm 7.8 (OPTICS: update seed list)

```
SeedList :: update(o,obj)
  compute reachdist(o,obj) =: current_rdist_o;
  IF o is already in SeedList THEN
    IF current_rdist_o < o.rdist THEN
      o.rdist := current_rdist_o;
      reorganize heap;
  ELSE
    insert o with o.rdist := current_rdist_o into SeedList;
```

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Intuitively: cluster ≈ valley)

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Cluster order as graph: jumps

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 2$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

("reached from" graph: $\approx$ spanning tree)

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Intuitively: cluster ≈ valley)
Previous "hill" would belong to the cluster, subsequent "hill" does not!

OPTICS – Example $\varepsilon = 5, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Plot is smoothed!

Heap happens to be more expensive here (not necessarily!)

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

(Intuitively: cluster ≈ valley)
Previous "hill" would belong to the cluster, subsequent "hill" does not!

OPTICS – Example $\varepsilon = 2, \mathrm{MinPts} = 4$

Start example | Skip example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

radius too small!

OPTICS Plot: Reachability Diagram

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

OPTICS Plot: Reachability Diagram

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

The cluster order (or: reachability diagram) shows the reachability distances w.r.t. ε and MinPts of the objects as bars the order (x-axis) given by the "cluster order"

cut: $\approx$ density-level as ε for DBSCAN

OPTICS Parameter Sensitivity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- cluster order relatively robust against parameter choices
- reasonable result when parameters are “big enough”

OPTICS Parameter Sensitivity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- cluster order relatively robust against parameter choices
- reasonable result when parameters are “big enough”

Note that:

The meaning and effect of ε is quite different from the meaning of the equally named parameter in DBSCAN.

Manual Data Analysis with OPTICS

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- are there clusters?
- how many?
- are the clusters hierarchically nested?
- how big are the clusters?

Cluster Extraction from Hierarchies

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Hierarchical DBSCAN* (HDBSCAN*) [Campello et al., 2013a, 2015]: formulates a criterion of cluster stability, optimizes overall quality of clusters extracted by *local* cuts
- framework for cluster extraction [Campello et al., 2013b]: finds the globally optimal *local* cut, based on additional criteria – HDBSCAN* becomes special case [Campello et al., 2015]

Videos

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Non-Parametric Density Estimation
Density-based Clustering
Hierarchical Clustering
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

A research group in Canada implemented HDBSCAN in scikit-learn [McInnes and Healy, 2017].
They talk about it in some conference presentations:

- *comparing discussion of various clustering methods and a demonstration of HDBSCAN**
  `https://www.youtube.com/watch?v=AgPQ76RIi6A`
- *from the same group, similar content, but a bit more context and technical explanation:*
  `https://www.youtube.com/watch?v=dGsxd67IFiU`

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Recommended Reading:

- *textbooks:*
  - *Tan et al. [2006], Chapter 10*
  - *Han et al. [2011], Chapter 12*
- *research articles:*
  - *General overview, interpretation, and categories: Zimek and Filzmoser [2018]*
  - *Advanced reading on locality and density-estimates: Schubert et al. [2014a], Schubert et al. [2014b].*
  - *On the problems of evaluation: Campos et al. [2016], Schubert et al. [2012].*
  - *Advanced reading on specializations to high-dimensional data: Zimek et al. [2012].*

Outline

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Probabilistic Outlier Model: Statistical Tests

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

general idea

- given a certain kind of statistical distribution (e.g., Gaussian)
- compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)
- outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean)

basic assumptions

- normal data objects follow a (known) distribution and occur in a high probability region of this model
- outliers deviate strongly from this distribution

Probabilistic Outlier Model: Statistical Tests

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Many statistical tests are available in the literature, see classic textbooks (e.g., by Barnett [1978] or by Hawkins [1980]). Such tests differ in assumptions on:

- the type of data distribution (e.g., Gaussian)
- the number of variables, i.e., dimensions of the data objects (univariate/multivariate)
- the number of distributions (mixture models)
- parametric versus non-parametric (e.g., histogram-based)

example on the following slides:

- Gaussian distribution
- multivariate
- 1 model
- parametric

Probabilistic Outlier Model: Statistical Tests

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

probability density function (multivariate normal distribution):

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)\cdot(\Sigma)^{-1}\cdot(x-\mu)^{\mathsf{T}}}$$

- μ is the mean value of all points (usually data are normalized such that $\mu = 0$)
- Σ is the covariance matrix centered at the mean
- MDist$= (x-\mu)\cdot(\Sigma)^{-1}\cdot(x-\mu)^{\mathsf{T}}$ is the Mahalanobis distance of point x to μ
- MDist follows a χ^2-distribution with d degrees of freedom ($d =$ data dimensionality)
- all points x, with MDist$(x,\mu) > \chi^2(0.975)$ $[\approx 3\cdot\sigma]$ are considered outliers

Visualization

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Example (2D)

(Figure from Tan et al. [2006].)

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Distance-based Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

DB(ε, π)-outlier [Knorr and Ng, 1997]

- given ε, π
- A point p is considered an outlier if at most π percent of all other points have a distance to p less than ε

$$OutlierSet(\varepsilon, \pi) = \left\{ p \,\middle|\, \frac{Cardinality(q \in \mathcal{D} \mid \mathrm{dist}(q,p) < \varepsilon)}{Cardinality(\mathcal{D})} \leq \pi \right\}$$

Distance-based Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Outlier scoring based on kNN distances:

- Take the kNN distance of a point as its outlier score [Ramaswamy et al., 2000]

- Variant: Aggregate the distances for the 1-NN, 2-NN, . . . , kNN (sum, average) [Angiulli and Pizzuti, 2002]

Choosing Parameters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

(Figures from Tan et al. [2006].)

Choosing Parameters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

(Figures from Tan et al. [2006].)

Choosing Parameters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

(Figures from Tan et al. [2006].)

Choosing Parameters

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

(Figures from Tan et al. [2006].)

Density-based Local Outliers

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Figure from Breunig et al. [2000].

- DB-outlier model: no parameters ε, π such that o_2 is an outlier but none of the points of C_1 is an outlier
- kNN-outlier model: kNN-distances of points in C_1 are larger than kNN-distances of o_2

Density-based Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Local Outlier Factor (LOF) [Breunig et al., 2000]:

- reachability distance (smoothing effect):
  $\mathrm{reachdist}_k(p,o) = \max\{k\,\mathrm{dist}(o), \mathrm{dist}(p,o)\}$
- local reachability density (lrd)
  $\mathrm{lrd}_k(p) = 1/\frac{\sum_{o \in kNN(p)} \mathrm{reachdist}_k(p,o)}{Cardinality(kNN(p))}$
- Local outlier factor (LOF) of point p: average ratio of lrds of neighbors of p and lrd of p

Figure from Breunig et al. [2000]

$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(p)}}{Cardinality(kNN(p))}$$

- LOF ≈ 1: homogeneous density
- LOF $\gg 1$: point is an outlier (meaning of "$\gg$" ?)

Distance-based vs. Density-based?

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Discussion:

- *Is there a fundamental difference between "distance-based" methods and "density-based" methods?*
- *What is the common underlying principle?*

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Distance to nearest neighbor

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Distance to nearest neighbor
  $\Rightarrow$ misses paired outliers

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Distance to k nearest neighbor

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Distance to k nearest neighbor

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Distance to k nearest neighbor
  $\Rightarrow$ micro clusters ($|C| < k + 1$) become outliers

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Sum (average) of distances to the first k nearest neighbors
  More robust with respect to k and micro clusters

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Sum (average) of distances to the first k nearest neighbors

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Distance based outliers:
"Outliers are farther away from the remainder of the data than most other objects."

- Sum (average) of distances to the first k nearest neighbors

Local Density Estimate, Global Comparison

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Note that:

- *Both*
  - *the k nearest neighbor distance*
  - *and the average k nearest neighbor distance*

  are variants of local density estimates.
- *All local density estimates are ranked, the lowest density estimate is the strongest outlier.*
- *We cannot handle areas of different densities, because all density estimates are compared globally, against all others.*

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Local outlier factor:

Idea:

Outliers exhibit a lower local density than their neighbors.

Global vs. Local Outliers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Local outlier factor:

Idea:

Outliers exhibit a lower local density than their neighbors.

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \frac{\sum_{o \in \mathcal{N}} \frac{lrd_k(o)}{lrd_k(p)}}{|\mathcal{N}|}$$

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \frac{1}{|\mathcal{N}|} \sum_{o \in \mathcal{N}} \frac{lrd_k(o)}{lrd_k(p)}$$

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \underbrace{\frac{1}{|\mathcal{N}|}\sum_{o\in\mathcal{N}}}_{\text{average}} \underbrace{\frac{lrd_k(o)}{lrd_k(p)}}_{\text{relative density}}$$

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \underbrace{\frac{1}{|\mathcal{N}|}\sum_{o\in\mathcal{N}}}_{\text{average}} \underbrace{\frac{lrd_k(o)}{lrd_k(p)}}_{\text{relative density}}$$

- Same density $\Leftrightarrow$ relative density $= 1$

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \underbrace{\frac{1}{|\mathcal{N}|}\sum_{o\in\mathcal{N}}}_{\text{average}} \underbrace{\frac{lrd_k(o)}{lrd_k(p)}}_{\text{relative density}}$$

- Same density $\Leftrightarrow$ relative density $= 1$
- Less dense $\Leftrightarrow$ relative density > 1

Local Outlier Factor

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

$$LOF(p) = \underbrace{\frac{1}{|\mathcal{N}|}\sum_{o\in\mathcal{N}}}_{\text{average}} \underbrace{\frac{lrd_k(o)}{lrd_k(p)}}_{\text{relative density}}$$

- Same density $\Leftrightarrow$ relative density $= 1$
- Less dense $\Leftrightarrow$ relative density > 1
- $LOF(p) \gg 1$ for outliers!

Generalized Locality in “Local Outliers”: Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
    select context(o);

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
    select context(o);

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in \mathcal{O}$ **do**

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in \mathcal{O}$ **do**
  select reference(o);

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in \mathcal{O}$ **do**
  select reference(o);

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in \mathcal{O}$ **do**
  select reference(o);

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in \mathcal{O}$ **do**
  select reference(o);
  score(o) := compare(model(o), $\{\text{model}(r)\}_{r \in \text{reference}(o)}$);
end

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database O

foreach $o \in O$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

foreach $o \in O$ **do**
  select reference(o);
  score(o) := compare(model(o), $\{\text{model}(r)\}_{r \in \text{reference}(o)}$);
end

foreach $o \in O$ **do**
  normalizedScore(o) := normalize(score(o));
end

Generalized Locality in "Local Outliers": Schema [Schubert et al., 2014a]

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Data: database $\mathcal{O}$

```
// STEP 1: model building
```

foreach $o \in \mathcal{O}$ **do**
  select context(o);
  model(o) := buildModel(o, context(o));
end

```
// STEP 2: model comparison
```

foreach $o \in \mathcal{O}$ **do**
  select reference(o);
  score(o) := compare(model(o), $\{\text{model}(r)\}_{r \in \text{reference}(o)}$);
end

```
// STEP 3: normalization
```

foreach $o \in \mathcal{O}$ **do**
  normalizedScore(o) := normalize(score(o));
end

Local vs. Global Outlier Model

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

general schema [Schubert et al., 2014a]

- context set for model building
- reference set for model comparison
- both, context and reference, could be:
  - identical
  - local
  - global

(Figure from Schubert et al. [2014a].)

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Variants of Outlier Models I

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- different notion of neighborhood:
  - connectivity-based (COF) [Tang et al., 2002]
  - reverse neighborhood:
    - INFLO [Jin et al., 2006]

    - ODIN [Hautamäki et al., 2004] (related to low hubness outlierness [Radovanović et al., 2014])

Variants of Outlier Models II

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- different comparison:
  - local outlier integral (LOCI) [Papadimitriou et al., 2003]
  - local distance-based outlier (LDOF) [Zhang et al., 2009]

- different density kernels:
  - LDF [Latecki et al., 2007]
  - KDEOS [Schubert et al., 2014b]
  - SimplifiedLOF [Schubert et al., 2014a]
  - LoOP [Kriegel et al., 2009]

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Outlier vs. Inlier – a Classification Problem?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- class imbalance: the “class” outlier is substantially smaller, but often more important than the “class” inlier
- resulting problems:
  - training of a classifier is difficult due to lack of training data
  - evaluation measures are hard to interpret (e.g., typically very low precision values even for a rather good result)
- most (unsupervised) outlier detection methods deliver not a decision but a score (“outlier score”, “outlier factor”)
- outlier detection result based on scores is a ranking

Evaluation of Rankings

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- methods return a full ranking of database objects
- user interested in the top-ranked objects

examples taken from Campello et al. [2015]

Evaluation of Rankings

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- two-class confusion matrix ($C(o)$: class, $P(o)$: prediction):

$C(o) \backslash P(o)$	*outlier*	*inlier*
outlier	*TP*	*FN*
inlier	*FP*	*TN*

- ranking is a continuum between strongest outlier and weakest outlier (i.e., strongest inlier): we consider only the first column (everything is predicted being an outlier, but in an order as defined by the ranking)
- ideal ranking: all TP before any FP

ROC: Receiver Operating Characteristic

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- each TP in the ranking: one step up
- each FP in the ranking: one step to the right
- comparison of two rankings: area under the curve (ROC AUC)

- $0 \leq$ ROC AUC ≤ 1
- interpretation: probability that two randomly chosen objects, one positive example (outlier) and one negative example (inlier), have a correct relative ranking (with the outlier ranked before the inlier) [Hanley and McNeil, 1982]

External vs. Internal Evaluation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- Evaluation with ROC needs class labels (external information).
- This kind of evaluation is therefore called *external evaluation*.
- In an application of an unsupervised method, we would actually not have external information available.
- We need *internal evaluation*, relying only on the data itself.

Internal Evaluation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Marques et al. [2015] proposed the first internal evaluation measure, based on a notion of classification hardness.

Internal Evaluation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Marques et al. [2015] proposed the first internal evaluation measure, based on a notion of classification hardness.

Internal Evaluation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Marques et al. [2015] proposed the first internal evaluation measure, based on a notion of classification hardness.

Internal Evaluation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Marques et al. [2015] proposed the first internal evaluation measure, based on a notion of classification hardness.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
  - Parametric Outlier Detection: Probabilistic Outlier Model
  - Non-Parametric Outlier Detection
  - Global vs. Local Outliers
  - Variants of Outlier Models
  - Evaluation of Outlier Detection Methods
  - Some Fundamental Problems
- Summary

Unavoidable False Positives and False Negatives

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- objects truly generated by a different mechanism can fit very well to the "normal" data distribution
- inliers in the tail of the "normal" distribution must appear suspicious

decision on true outlierness based on domain knowledge

Masking and Swamping

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- mean and standard deviation are sensitive to outliers themselves
- these values are computed for the complete data (including potential outliers) – masking and swamping:

- the MDist is used to determine outliers although the MDist values are influenced by these outliers (advanced approaches try to alleviate this effect)
- applies to any model (some are more robust, some are more susceptible to these effects)
- thus statistical research on “robust methods” [Rousseeuw and Hubert, 2011]

Masking and Swamping

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- mean and standard deviation are sensitive to outliers themselves
- these values are computed for the complete data (including potential outliers) – masking and swamping:

- the MDist is used to determine outliers although the MDist values are influenced by these outliers (advanced approaches try to alleviate this effect)
- applies to any model (some are more robust, some are more susceptible to these effects)
- thus statistical research on “robust methods” [Rousseeuw and Hubert, 2011]

Masking and Swamping

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- mean and standard deviation are sensitive to outliers themselves
- these values are computed for the complete data (including potential outliers) – masking and swamping:

- the MDist is used to determine outliers although the MDist values are influenced by these outliers (advanced approaches try to alleviate this effect)
- applies to any model (some are more robust, some are more susceptible to these effects)
- thus statistical research on “robust methods” [Rousseeuw and Hubert, 2011]

Curse of Dimensionality

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Parametric
Non-Parametric
Global vs. Local
Variants
Evaluation
Problems
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

- curse of dimensionality: the larger the degree of freedom, the more similar the MDist values are for all points

x-axis: observed *MDist* values

y-axis: frequency of observation

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
- Discrete Random Variables and Expectation
- Moments and Deviations
- Continuous Distributions
- Bayesian Learning with Distributions (Parametric Learning)
- Non-Parametric Learning
- Outlier Detection — Parametric and Non-Parametric
- Summary

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *discrete random variables and some discrete distributions with their parameters*
- *expectation, variance, standard deviation, covariance, correlation*
- *continuous distributions (uniform and normal) and their parameters*
- *Bayesian learning with distributions*
- *algebraic vs. probabilistic view on data*
  - *EM clustering and a probabilistic interpretation of k-means*
  - *LDA*

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *parametric vs. non parametric learning*
- *non-parametric density estimation*
- *basic ideas of density-based clustering and the algorithm DBSCAN*
  - *core point*
  - *(direct) density-reachability*
  - *density-connectivity*
  - *border points*
  - *noise*
- *variants*
  - *include border points?*
  - *SNN-based clustering*
  - *GDBSCAN*

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *dendrogram as representation of a hierarchical clustering solution*
- *distances for clusters and objects and the basic agglomerative greedy algorithm*
- *density-based hierarchical clustering*
  - *formalization*
  - *algorithm OPTICS*
  - *cluster order*
    - *construction*
    - *interpretation*
    - *parameters*

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Discrete Random Variables and Exp.
Moments and Deviations
Continuous Distr.
Bayesian Learning II
Non-Parametric Learning
Outlier Detection
Summary
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
References

You learned in this section:

- *probabilistic parametric model for outlier detection*
- *non-parametric models:*
  - *distance-based models*
    - *DB-outlier model*
    - *kNN-based model*
  - *density-based model*
    - *LOF: motivation, model*
    - *pointers to the literature: many variants of LOF*
- *evaluation of outlier detection*
- *problems and challenges*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Recommended Reading:

- *Mitzenmacher and Upfal [2017], Chapter 10.*

Entropy Function

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Definition 8.1 (Entropy of a Discrete Random Variable)

The *entropy* in bits of a discrete random variable X is given by

$$H(X) = -\sum_x \Pr(X = x) \log_2 \Pr(X = x),$$

where the summation is over all values x in the range of X.

Corollary 8.2

$$H(X) = E\left[\log_2 \frac{1}{\Pr(X)}\right]$$

Binary Entropy Function

Definition 8.3 (Binary Entropy Function)

The *binary entropy function* of a random variable that assumes only two possible outcomes occurring with probability p and $1-p$, respectively, is

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

We define $H(0) = H(1) = 0$.

n-ary Entropy (Uniform)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Generalizing, the entropy of a random variable X with n outcomes of equal probability is given by:

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} \\
&= -\sum_{i=1}^{n} \log_2 \left(\frac{1}{n}\right)^{\frac{1}{n}} \\
&= -\log_2 \left(\left(\frac{1}{n}\right)^{\frac{1}{n}}\right)^{n} \\
&= -\log_2 \frac{1}{n} \\
&= \log_2 n
\end{aligned}
$$

Intuition

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- Imagine a die with 8 sides that show up with equal probability.
- The entropy is 3 bits.
- If the faces of the die were numbered with 0 to 7 in binary code, the outcome of a die roll would give a sequence of 3 bits uniform over the set $\{0,1\}^3$.
- This shows the equivalence to generating 3 bits independently and uniformly at random.

Note that:

The entropy of a random variable X does not depend on the values that X can take but only on the probability distribution of X over those values.

Interpretations

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- One interpretation of entropy is that it measures the amount of randomness (disorder, uncertainty) in a system:
  - For example consider the second law of thermodynamics: the total entropy of an isolated system cannot decrease over time.
- Another interpretation relates entropy to compression and coding theory (see Slide 301):
  - Entropy relates to the number of bits per symbol required to encode a message.
  - A very predictable message can be encoded with less bits (it conveys less information).
  - An unpredictable message conveys much information (high entropy).

Entropy of Class Distributions

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

$$H(\mathcal{D}) = -\sum_{i=1}^{k} \Pr(c_i|\,\mathcal{D}) \log_2 \Pr(c_i|\,\mathcal{D})$$

- Considering labeled data with $k = 2$, a very pure set (almost all labels belong to class A, only some belong to class B) has very low entropy (i.e., low disorder, low uncertainty):

- If we draw some data object at random, we are very likely to get one that belongs to class A.
- The more equal the proportions of the two classes are, the more uncertainty we have about which class we will likely draw (i.e., higher disorder, higher entropy).

Gini Index

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Definition 8.4 (Gini Index)

The Gini index as a measure for the (im-)purity of a data set $\mathcal{D}$ w.r.t. k classes $c_1, \ldots, c_k$ is given by:

$$G(\mathcal{D}) = 1 - \sum_{i=1}^{k} \Pr(c_i | \mathcal{D})^2$$

- If a dataset contains only one class, the probability of that class is 1, the dataset has minimal impurity, the Gini index is 0.
- When each class is equally represented, we have $\Pr(c_i | \mathcal{D}) = \frac{1}{k}$, the dataset is maximally impure, and the Gini index is $\frac{k-1}{k}$ (i.e., approaching 1 as $k \to \infty$).
- Probabilistic interpretation of the square: if we randomly draw two objects from $\mathcal{D}$, how likely are they belonging to the same class?

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Recommended Reading:

text books

- *Zaki and Meira Jr. [2014], Chapter 19.*
- *Mitchell [1997], Chapter 3.*
- *Tan et al. [2006], Chapter 4.3.*
- *Witten et al. [2011], Chapter 6.1.*

fundamental paper

- *Quinlan [1986]*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Decision Tree – Basic Idea

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Data:

ID	age	car type	risk
1	23	family car	high
2	17	sports car	high
3	43	sports car	high
4	68	family car	low
5	32	truck	low

Decision tree:

- A decision tree provides explicit knowledge on the data.
- The classification model is interpretable (a hierarchy of rules).

Decision Tree – Properties

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- A decision tree is a tree with the following properties:
  - an inner node represents a test on an attribute
  - an edge represents a test result on the parent node
  - a leave node represents one of the classes

- Construction: top-down based on the training set
- Application (prediction):
  - traversal according to the tests from the root to some leave node (deterministic path)
  - class assignment: the class of the leave node reached in the traversal

Decision Tree – Basic Algorithm

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Algorithm 8.1 (Decision Tree)

1. *given a dataset, select an attribute and split point (greedy selection, following some split strategy)*
2. *partition the data according to the test on the split attribute*
3. *repeat the procedure recursively for each partition*

The recursion stops if the partition is "pure" (contains only examples of a single class).

Example: Should We Play Tennis Today?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

ID	forecast	temperature	humidity	wind	play tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example Decision Tree

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Types of Splits

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

categorical attributes

- attribute $= a$
- attribute $\in A$

numerical attributes

- attribute $< a$, $\leq a$
- atribute $\geq a$, $> a$

Problem: Where to Split Numerical Attributes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Where should we define a split for some numerical attribute?

- We want to maximize the separation of classes.
- Idea: sort attribute values

value	0.9	0.8	0.65	0.5	0.45	0.3	0.15	0.01
class	A	A	B	B	B	A	A	A

potential split points

- test combinations with split criterion

alternative:

- fit to each class a Gaussian distribution
- intersections of the Gaussian pdfs are potential split points

Quality Measure for Splits

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

To select attributes for splitting and to select split points in attributes, we need to assess the quality of partitions induced by a split on some attribute.

given

- a training set TR
- disjunct and complete partitionings $T = T_1, \ldots, T_m$ of TR: $\cup_i T_i = TR$, $\forall i \neq j : T_i \cap T_j = \emptyset$
- relative frequency of each class c_i in each partition T_j: $p_i = \Pr(c_i | T_j)$

required

- a relative measure for the purity w.r.t. classes of some set of partitions
- a split that optimizes this measure

Measure: Information Gain

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- Information gain is a measure based on the entropy.

$$H(T) = -\sum_{i=1}^{k} \Pr(c_i|T) \log_2 \Pr(c_i|T)$$

- Information gain measures the reduction of entropy (i.e., gain of information) by a split of set T into partitions $T_1, \ldots, T_m$:

$$\textit{information gain}(T, T_1, \ldots, T_m) = H(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} H(T_i)$$

- Higher information gain means larger reduction of entropy.
- We choose the attribute and split point that maximize the information gain.

Example: Should We Play Tennis Today?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

ID	forecast	temperature	humidity	wind	play tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example: Information Gain

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

9 “yes”, 5 “no”: $H(T) = 0.940$

humidity

high: 3 “yes”, 4 “no”: $H(T_1) = 0.985$

normal: 6 “yes”, 1 “no”: $H(T_2) = 0.592$

$$\textit{inf. gain}(T, T_i(\textit{humidity})) = 0.940 - \frac{7}{14}0.985 - \frac{7}{14}0.592 = 0.151$$

wind

weak: 6 “yes”, 2 “no”: $H(T_1) = 0.811$

strong: 3 “yes”, 3 “no”: $H(T_2) = 1.0$

$$\textit{inf. gain}(T, T_i(\textit{wind})) = 0.940 - \frac{8}{14}0.811 - \frac{6}{14}1.0 = 0.048$$

Measure: Gini Index

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

$$G(\mathcal{D}) = 1 - \sum_{i=1}^{k} \Pr(c_i | \mathcal{D})^2$$

In an analogous way, we can use the weighted Gini index of induced partitions to compare partitionings:

$$G(T_1, \ldots, T_m) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} G(T_i)$$

- Smaller value of the Gini index means lower impurity.
- We choose the attribute and the split that minimizes the Gini index.

Example: Should We Play Tennis Today?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

ID	forecast	temperature	humidity	wind	play tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example: Gini Index

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

9 “yes”, 5 “no”: $|T| = 14$

$$G(T_1) = 1 - \left(\frac{3^2}{7^2} + \frac{4^2}{7^2}\right) = \frac{24}{49} \qquad G(T_2) = 1 - \left(\frac{6^2}{7^2} + \frac{1^2}{7^2}\right) = \frac{12}{49}$$

$$G(T_1) = 1 - \left(\frac{6^2}{8^2} + \frac{2^2}{8^2}\right) = \frac{24}{64} = \frac{3}{8} \qquad G(T_2) = 1 - \left(\frac{3^2}{6^2} + \frac{3^2}{6^2}\right) = \frac{18}{36} = \frac{1}{2}$$

$$G(\textit{split on humidity}) = \frac{7}{14} \cdot \frac{24}{49} + \frac{7}{14} \cdot \frac{12}{49} = \frac{18}{49}$$

$$G(\textit{split on wind}) = \frac{8}{14} \cdot \frac{3}{8} + \frac{6}{14} \cdot \frac{1}{2} = \frac{3}{7} = \frac{21}{49}$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Axis-parallel Hyperplanes

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- A hyperplane $h(x)$ is defined as the set of all points $x \in \mathbb{R}^d$ that satisfy:

$$h(x) : w \cdot x^{\mathsf{T}} + b = 0,$$

  where w is a weight vector that is normal to the hyperplane, and b is the offset of the hyperplane from the origin.
- For axis-parallel hyperplanes, the weight vector is parallel to one of the axes, i.e., $w \in \{e_1, \ldots, e_d\}$, where $e_i \in \mathbb{R}^d$ has value 1 in dimension X_i and value 0 in every other dimension.

$$\begin{aligned} h(x) &: e_i \cdot x^{\mathsf{T}} + b = 0 \\ \equiv h(x) &: x_i + b = 0 \end{aligned}$$

- The choice of b yields different hyperplanes parallel to axis X_i.

Hyperplanes and Split Points

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- For decision trees on real-valued attributes, axis-parallel hyperplanes relate to split points.
- A hyperplane splits the data space $\mathbb{R}^d$ into two half-spaces.
- All points with $h(x) < 0$ are on one side of the hyperplane, all points with $h(x) > 0$ are on the other side of the hyperplane.
- We can therefore write the split point as:

$$h(x) \leq 0 \quad \Leftrightarrow \quad x_i + b \leq 0 \quad \Leftrightarrow \quad x_i \leq -b$$

- As b can be choosen to be any value of the dimension X_i, the generic form of a split point (or test) is

$$X_i \leq v,$$

where $v = -b$ is some value in the domain of X_i.

Bias: Axis-Parallel Piecewise Linear Separation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The bias of a decision tree is therefore to separate data along piecewise axis-parallel hyperplanes.

Figures from Zaki and Meira Jr. [2014].

Bias Prevents Detection of Certain Simple Decision Rules

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Consider a dataset with a relatively simple decision rule:

- if $x + y \leq 1 \Rightarrow +$
- else $\Rightarrow \bullet$

- A decision tree is biased to find a more complex model, as it can only find piecewise axis-parallel hyperplanes as split points.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Overfitting: Training Error vs. Test Error

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- Given two classifiers E and E'.
- Let E have smaller error on training data than E'.
- Let E' have smaller error on test data than E.
- Then E is overfitted on the training data in comparison with E'.

Overfitting

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- Decision trees (basic algorithm 8.1) are susceptible to overfitting in general. (Why? Cf. slide 215.)
- With increasing size, decision trees tend to overfit more.

Overfitting Scenario

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Training data with * mislabeled examples:

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	no*
whale	warm-blooded	yes	no	no	no*
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

Overfitting Scenario

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Test data:

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
human	warm-blooded	yes	no	no	yes
pigeon	warm-blooded	no	no	no	no
elephant	warm-blooded	yes	yes	no	yes
leopard shark	cold-blooded	yes	no	no	no
turtle	cold-blooded	no	yes	no	no
penguin	warm-blooded	no	no	no	no
eel	cold-blooded	no	no	no	no
dolphin	warm-blooded	yes	no	no	yes
spiny anteater	warm-blooded	no	yes	yes	yes
gila monster	cold-blooded	no	yes	yes	no

Two Models

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

(a) Model M1 (b) Model M2

training error: 0%
test error: 30%

training error: 20%
test error: 10%

M1:

- misclassification of dolphin and human due to overfitting (on mislabeled training data)
- spiny anteater: unusual case

M2:

- misclassification of unusual case cannot be avoided

Lack of Training Data

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Consider this small training set, containing unusual cases:

Name	Body Temp	Gives Birth	Four-legged	Hibernates	Mammal
salamander	cold-blooded	no	yes	yes	no
poorwill	warm-blooded	no	no	yes	no
platypus	warm-blooded	no	yes	yes	yes
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

The resulting decision tree has a test error of 30%:

Heuristics for Avoiding Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- remove erroneous training data
  (in particular contradictory training data)
- suitable amount of training data
  (training set should be large enough, but too large can be counterproductive as well, as it might contain too many special cases — it should be *representative*)
- choose a minimum support $\gg 1$ for leaves: number of training examples that need to belong to a leave node
- choose a minimum confidence $\ll 100\%$ (purity), as fraction that the majority class in a leave node needs to satisfy
  (leaves can absorb erroneous or unusual data, noise)

Error-Reduction-Pruning

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Cut overspecialized branches of the tree (as smaller trees are less susceptible to overfit – cf. Occam's razor, slide 299):

Algorithm 8.2 (Error-Reduction-Pruning)

- *Split training data into model training data MTR and model selection data MS.*
- *Construct decision tree E on MTR.*
- *Prune E on MS:*
  - *Find the subtree E_i such that reduction of the error on MS is maximal for $E \setminus E_i$.*
  - *Remove E_i from E.*
  - *Stop if no such subtree exists.*

General Principle: Model Tuning and Data Snooping

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- Assume we want to tune some model, e.g., by optimizing some parameter:
  - Train the classifier with some parameter setting on the training data TR.
  - Test the performance of the learned model with the given parameter on the test set TE.
  - Repeat the procedure for a different parameter setting (possibly using grid search for the optimal setting).
- In such a setting, we made use of the test set TE for optimizing the model: we ask new questions to the training data, based on the answers we got from the test data.
- An additional test set is required, that is independent of both, the training set TR and the test set TE.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
  - Basic Algorithm
  - Splits and Split Criteria
  - Geometric Interpretation and Bias
  - Overfitting and Error-Reduction Pruning
  - Discussion
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Basic Algorithm
Splits/Split-Crit.
Geom. Interp./Bias
Overfitting/Error-Red. Pruning
Discussion
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

pro

- The decision tree model is easily interpretable and provides human-readable information on the data.
- The structure of the tree provides an implicit weighting of attributes.
- Decision trees are typically strong classifiers and are often used in practice.
- The classifier facilitates efficient application of the derived classification model on new data.

con

- Finding the optimal tree is exponential.
- Heuristic (greedy!) tree building algorithms can only find local optimum.
- Decision trees are susceptible to overfitting (error reduction pruning should be considered).

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
  - Basic Functionality
  - Geometrical Interpretation: Linear vs. Non-Linear Separation
  - Learning Algorithm and Network Design
  - Discussion
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Recommended Reading:

- *Mitchell [1997], Chapter 4*
- *Tan et al. [2006], Chapter 5.4*

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
  - Basic Functionality
  - Geometrical Interpretation: Linear vs. Non-Linear Separation
  - Learning Algorithm and Network Design
  - Discussion
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Motivation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

- Neural networks have been developed as a machine and computational model, simulating some principles of biological brains:
  - neural network (or short: neural net) is a set of neurons that are connected by edges
  - neuron: mimics the function of a biological neuron, being activated by input-signals at synapses and firing impulses to other neurons under certain conditions
- Human brain:
  - 10^{11} neurons each connected with approx. 10^4 others
  - fastest response time: 10^{-3} sec (3GHz processor: approx. 10^{-10} sec.)
  - derives complex decisions very fast (recognizing the face of your mother typically takes 100 steps, i.e., 10^{-1} sec.)

Basic Functionality

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

- conditional creation of output signal that is transmitted to other neurons
- simplified construction of the artificial neural net:
  - input layer
  - hidden layers
  - output layer
  - node in some layer is connected with all nodes of the previous layer
  - connections are directed edges with weights
  - the weights are adapted to model a given learning problem
- the architecture of biological brains is more complex (in most animals) and variable

Architecture

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

output vector y

output layer

hidden layer

input layer

input vector x

Neurons

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

general neuron (a.k.a. perceptron):

activation value a

$$a = \sum_{i=1}^{n} w_i \cdot x_i$$

$$y = \frac{1}{1+e^{-a}}$$

threshold logic unit (TLU):

all or nothing,
threshold θ

$$y = \begin{cases} 1 & \text{if } a \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Example: Model for a Boolean Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

$$f(x) = [x_1 \wedge (x_2 \vee x_3)] \vee (x_2 \wedge x_3)$$

x_1	x_2	x_3	y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0

To build the classifier, we need to learn the weights w_1, w_2, w_3.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
  - Basic Functionality
  - Geometrical Interpretation: Linear vs. Non-Linear Separation
  - Learning Algorithm and Network Design
  - Discussion
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Geometrical Interpretation of a TLU Model

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

- $\sum_{i=1}^{d} w_i \cdot x_i = \theta$: hyperplane in d-dimensional space
- TLU learnes a separating hyperplane: examples of class 0 should be on one side, examples of class 1 on the other.

- The training of a TLU is about adapting iteratively the weights w_i until all training examples are separated by the hyperplane according to their respective class.
- Training step: rotation of the hyperplane defined by w_i and θ in the direction of point v if v is not (yet) on the correct side of the hyperplane.

Not Linearly Separable Learning Problem

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

The XOR function, represented in a two dimensional space $[0, 1]^2$, is not linearly separable:

X_1	X_2	y
0	0	−1
1	0	1
0	1	1
1	1	−1

Combination of Several Neurons

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Non-linearly separable classes require a hidden layer:

Architecture variants: feed-forward or recurrent (allows also connections within the same layer or to previous layers)

Solution for XOR

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

A possible neural net modelling the XOR function:

Geometrical Interpretation of the Neural Net Model

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

A hidden layer models the combination of several hyperplanes:

$$h_1 = 0 \wedge h_2 = 0 : y = 0 \text{ (classB)}$$

$$\text{otherwise} : y = 1 \text{ (classA)}$$

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
  - Basic Functionality
  - Geometrical Interpretation: Linear vs. Non-Linear Separation
  - Learning Algorithm and Network Design
  - Discussion
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Learning Algorithm for Complex Networks

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

As long as the predicted class is not the given class on the examples: adaptation of the weights for several nodes.

- gradient descent (optimization approach) to minimize the overall error (backpropagation algorithm)
- Overall error is the sum of quadratic deviations of the output (vector) y from the desired output for the training data (least squares optimization).
- Required: output y is a *continuous and differentiable* function of the activation a.

Design of the Network Topology

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Decisions on

- number of input nodes
- number of hidden layers and number of their nodes, respectively
- number of output nodes

have strong input on generalization ability:

- not enough nodes: poor classification accuracy
- too many nodes: overfitting

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
  - Basic Functionality
  - Geometrical Interpretation: Linear vs. Non-Linear Separation
  - Learning Algorithm and Network Design
  - Discussion
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
Basic Functionality
Separation
Learning Algorithm and Network Design
Discussion
SVM
Regression
Summary
Ensemble Learning
References

pros

- typically good classification accuracy achievable, arbitrarily complex decision boundaries (choose a topology as simple as possible to avoid overfitting)
- redudant features are not problematic: weights will become small
- application is efficient

cons

- model (of more complex networks with hidden layers) is hardly interpretable: learns only weights, no class description
- inefficient learning (training can take a tremendous amount of time)
- susceptible to errors in training data
- possibly converging towards a *local* minimum

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
  - Optimal Linear Separation
  - Non-Linearly Separable Problems
  - Kernels
  - Discussion
- Regression
- Summary

Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Recommended Reading:

- *Zaki and Meira Jr. [2014], Chapter 21.*
- *Tan et al. [2006], Chapter 5.5.*
- *Advanced reading: Cristianini and Shawe-Taylor [2000], Schölkopf and Smola [2002].*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
  - Optimal Linear Separation
  - Non-Linearly Separable Problems
  - Kernels
  - Discussion
- Regression
- Summary

Ensemble Learning

Linear Separation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

- Given a classification problem in $\mathbb{R}^d$.
- Two classes are linearly separable, i.e., linear separation of the examples of two classes by a hyperplane is possible.
- There are still infinitely many possible separating hyperplanes – which one should we choose?

Which is the Best Separating Hyperplane?

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Criteria for selecting a good separating hyperplane:

- stability
- distance to objects of both classes

Maximum Margin Hyperplane

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

- maximal distance to objects of both classes (minimum ξ)
- minimal probability that hyperplane has to be adapted when new training examples are added
- best generalization

The maximum margin hyperplane (MMH) is *maximally stable*.

Note that:

The definition of the MMH depends only on those objects p_i that have a distance ξ to the hyperplane. Those points p_i are called support vectors.

Optimization Problem

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Hyperplane $H(x) : w^{\mathsf{T}} \cdot x + b = \langle w, x \rangle + b = 0$

Distance $\mathrm{dist}(p, H(x)) = \left| \frac{1}{\sqrt{\langle w,w \rangle}} \langle w, p \rangle + b \right|$

condition 1: no classification error

let $f(p_i) = A : y_i = -1, f(p_i) = B : y_i = 1$

$$\left.\begin{array}{lll} (y_i = -1) & \Rightarrow & \langle w, p_i \rangle + b < 0 \\ (y_i = 1) & \Rightarrow & \langle w, p_i \rangle + b > 0 \end{array}\right\} \Leftrightarrow y_i \cdot [\langle w, p_i \rangle + b] > 0$$

condition 2: maximal margin maximize ξ, where

$$\xi = \min_{p_i \in TR} \left| \frac{1}{\sqrt{\langle w, w \rangle}} \langle w, p_i \rangle + b \right|$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
  - Optimal Linear Separation
  - Non-Linearly Separable Problems
  - Kernels
  - Discussion
- Regression
- Summary

Ensemble Learning

Soft Margin

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

If the data are
not linearly separable

or the separation is not very stable

we can optimize with a trade-off between training error and size of the margin.

Soft Margin

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Consider training error during optimization:

- ξ_i: distance of p_i from the border of the margin (slack variable)
- regularization of the influence of each individual observation

Transform Featurespace

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

- If the problem is not linearly separable in the given feature space, the problem can be transformed into another feature space.
- This means, the hypothesis space is extended.

Feature Transformation – Principle

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Try to find a linear separation in the transformed feature space.

Example transformation:

A hyperplane in this feature space corresponds to a polynomial (2nd degree) separation surface in the original feature space.

Feature Transformation – Example

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Feature space $x = (x_1, x_2) \in \mathbb{R}^2$

Transformed feature space e.g.,

$$\Phi(x) = (x_1^2, x_2^2, \sqrt{2} \cdot x_1, \sqrt{2} \cdot x_2, \sqrt{2} \cdot x_1 \cdot x_2, 1) \in \mathbb{R}^6$$

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
  - Optimal Linear Separation
  - Non-Linearly Separable Problems
  - Kernels
  - Discussion
- Regression
- Summary

Ensemble Learning

Kernel Trick

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

For the optimization, we need to eventually compute scalar products between points.

A special similarity function, a *kernel*, is used for an implicite feature transformation:

- Instead of computing new features for all objects: $\Phi(x)$, and then compute scalar products in the new space,
- a kernel computes the scalar product in the transformed space directly, without explicitly computing the transformed space first:

$$K_\Phi(p_i, p_j) = \langle \Phi(p_i), \Phi(p_j) \rangle$$

- This is an interesting property for computationally expensive feature transformations ("kernel trick").

Kernel

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

When is a function a kernel function?

Theorem 8.5 (Mercer's Theorem)

A function $K : \mathbb{R}^n \to \mathbb{R}_0^+$ is a kernel function, if the kernel-matrix (Gram matrix)

$$KM(K) = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}$$

is positive semi-definite (i.e., the Gram matrix does not have any negative Eigenvalues).

Necessary Conditions for a Kernel Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Necessary (but not sufficient) conditions for K being a kernel function are:

symmetry

$$\begin{aligned} K_\Phi(p_i, p_j) &= \langle \Phi(p_i), \Phi(p_j) \rangle \\ &= \langle \Phi(p_j), \Phi(p_i) \rangle \\ &= K_\Phi(p_j, p_i) \end{aligned}$$

Cauchy-Schwarz $K_\Phi(p_i, p_j)^2 \leq K_\Phi(p_i, p_i) \cdot K_\Phi(p_j, p_j)$

Combinations of Kernels

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

For given Kernels K_1, K_2, a constant $a \in \mathbb{R}^+$, and a symmetric, positive semi-definite matrix B, the following functions K are also kernels:

- $K(x, y) = K_1(x, y) \cdot K_2(x, y)$
- $K(x, y) = K_1(x, y) + K_2(x, y)$
- $K(x, y) = a \cdot K_1(x, y)$
- $K(x, y) = x^\mathsf{T} \cdot B \cdot y$

Typical Kernels

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Typical examples for kernel functions are:

linear $K(x, y) = \langle x, y \rangle$

polynomial $K(x, y) = (\langle x, y \rangle + c)^d$

radial basis function $K(x, y) = e^{-\gamma \cdot |x-y|^2}$

Gauss kernel $K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

Sigmoid $K(x, y) = \tanh(\gamma \cdot \langle x, y \rangle + c)$

Decision Boundaries

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Radial basis function

Polynomial (2nd degree)

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
  - Optimal Linear Separation
  - Non-Linearly Separable Problems
  - Kernels
  - Discussion
- Regression
- Summary

Ensemble Learning

Discussion

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Optimal Linear Separation
Non-Linearly Sep. Problems
Kernels
Discussion
Regression
Summary
Ensemble Learning
References

pros

- can generate classifiers of high accuracy
- can be formulated as a convex optimization problem, thus efficient optimization algorithms can find the *global* optimum
- relatively low tendency to overfit (based on generalization theory) – but one should prefer simpler kernels
- efficient application of the learned model
- compact models

cons

- often extensive training time
- non-trivial implementation (note that we did not go into details of the optimization algorithms)
- derived classification models are hard to interpret

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Recommended Reading:

- *Tan et al. [2006], Appendix D.*

Regression vs. Classification

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

- In the classification problem, each object belongs to one of a set of (discrete) classes $c_i \in C = \{c_1, \ldots c_k\}$.
- The function we want to learn or approximate is a function $f : D \to C$, i.e., $f(x) \in C$.
- In the regression problem, the prediction aims at a real value.
- To each object x belongs a unique value in $\mathbb{R}$, i.e., the function we want to predict or approximate is a function $f : D \to \mathbb{R}$.
- examples:
  - prediction of selling rates of some product
  - recommended amount of fertilizer for a given type of soil
- The predicted value $f(x)$ is also called *predicted variable*, *response variable*, or *dependent variable* and typically denoted by y.

Different Regression Tasks

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

linear regression We assume that the predicted variable y follows a linear function.

multiple regression The input variable x is multi-dimensional (a vector).

non-linear regression General case – the regression function does not need to be linear.

Linear Regression

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Given Objects x_i, sampled from random variable X; s training objects are additionally annotated with a value following the random variable Y.

Assumption $Y = \alpha + \beta \cdot X$
The observed values y deviate from a line with constant variance (distances ε_i of y_i at x_i from the line).

Objective Find a line describing the expected values of $(Y|X)$.

Solution Method of least squares: determine α and β such that $L = \sum_{i=1}^{s} \varepsilon^2 = \sum_{i=1}^{s} (y_i - \alpha - \beta \cdot x_i)^2$ is minimal.

Solution

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Partial derivations return estimates for the line's

slope

$$\beta = \frac{\sum_{i=1}^{s}(x_i - E(X))(y_i - E(Y))}{\sum_{i=1}^{s}(x_i - E(X))^2}$$

Y offset

$$\alpha = E(Y) - \beta \cdot E(X)$$

Multiple Regression

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

In multiple (linear) regression we are seeking a regression hyperplane.

Given Objects $x \in \mathbb{R}^d$

Assumption Y is dependent on a linear combination of the d attributes of X:

$$Y = \alpha + \sum_{k=1}^{d} \beta_k \cdot X_k$$

Solution Minimize quadratic error

$$L = \sum_{i=1}^{s} \varepsilon^2 = \sum_{i=1}^{s} \left(y_i - \alpha - \sum_{k=1}^{d} (\beta_k \cdot x_{i,k}) \right)^2$$

Non-linear Regression

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

Given random variables X and Y

Assumption Y depends on X non-linearly

Example We assume polynomial dependency:

$$Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$$

Solution Define new variables $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$.

Solve multiple linear regression

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Overfitting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

The higher the degree of the polynomial, the more degrees of freedom the approximation has, the better can the observations be approximated.

- outliers might influence the model strongly
- overfitting problem

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
- Entropy, Randomness, and Information
- Decision Tree Learning
- Neural Networks
- Support Vector Machines
- Regression
- Summary

Ensemble Learning

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Entropy
Decision Tree Learning
Neural Networks
SVM
Regression
Summary
Ensemble Learning
References

You learned in this section:

- *Entropy and Gini-index as measures of randomness and purity*
- *Decision tree learning*
  - *Basic greedy algorithm for decision tree learning*
  - *strategies for selection of attributes and of split points*
  - *bias and geometrical interpretation (separating piece-wise axis-parallel hyperplanes)*
  - *overfitting and error-reduction pruning (model selection)*
- *Linear vs. non-linear separation and learning*
  - *Basics on Artificial Neural Nets*
  - *Basics on Support Vector Machines*
  - *Basics on regression*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
- Summary

Recommended Reading

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

Recommended Reading:

textbooks

- *Tan et al. [2006], Chapter 5.6*
- *Zaki and Meira Jr. [2014], Chapter 22*
- *Witten et al. [2011], Chapter 8*
- *Shalev-Shwartz and Ben-David [2014], Chapter 5 (advanced)*

papers

- *Dietterich [2000]*
- *Dietterich [2003]*
- *Valentini and Masulli [2002]*

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
- Summary

Classification as Approximation of the Target Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

Recall the classification task (slide 198f.):

- For some domain $\mathcal{D}$ and a set of classes $C = \{c_1, \ldots, c_k\}$, $k \geq 2$, each $o \in \mathcal{D}$ belongs uniquely to some $c_i \in C$, i.e., there is a function $f : \mathcal{D} \to C$.
- Given a set of objects $O = \{o_1, o_2, \ldots, o_n\} \subseteq \mathcal{D}$ and a mapping $(O \to C) \subset f$ (examples):
  We want to also map any object $o_m \in \mathcal{D} \setminus O$ to C.
- A classifier is trained on some training set $TR \subseteq O$ to learn the mapping function (a model or hypothesis) $h : \mathcal{D} \to C$.
- Ideally we have $\forall o \in TR : h(o) = f(o)$ (if not for all, we should have this at least for most examples o).
- In general, the hypothesis h should be an approximation of f.

Hypothesis Space and Target Function

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

- The true function f (target function) is unknown.
- Based on the training data and its hypothesis space $\mathcal{H}$, a learning algorithm looks for the hypothesis $h_i \in \mathcal{H}$ that fits optimally to the training data.

- The true function f is not necessarily an element of the hypothesis space!

Approximation and Accuracy

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

- We apply h on elements $x \in \mathcal{D}$ to predict class $c_i = f(x)$.
- The accuracy of a classifier (hypothesis h) is the probability (statistically/empirically: frequency) of its predictions being correct:

$$\mathrm{acc}(h) = \Pr(h(x) = f(x))$$

or

$$\mathrm{err}(h) = \Pr(h(x) \neq f(x)) = 1 - \mathrm{acc}(h)$$

(compare slide 226)

Reduction of Error Probability by Averaging

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

- The basic idea of using ensembles (combinations of several classifiers) is the reduction of the probability of errors by asking a jury of experts instead of just one expert and by letting them vote to find a common prediction.
- Intuitively, we expect a better approximation of f by aggregating (e.g., averaging) over several approximations h_i.

- Cf. the considerations on Bayes optimal classification (slide 305ff.).

Simple Voting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

- If we have two classes $C = \{-1, 1\}$, a simple voting procedure could be defined as:
  - Learn hypotheses $h_1, \ldots, h_k$ with associated weights $w_1, \ldots, w_k$.
  - ensemble classifier $\hat{h}$ is given by:

$$\hat{h}(x) = \begin{cases} \textit{if } w_1 h_1(x) + \ldots + w_k h_k(x) \geq 0 & : x \mapsto 1 \\ \textit{if } w_1 h_1(x) + \ldots + w_k h_k(x) < 0 & : x \mapsto -1 \end{cases}$$

- We can have $w_1 = \ldots = w_k = 1$ (i.e., unweighted voting).
- Weights can be based on (empirical) reliability of individual classifiers.
- We can have more complex voting procedures (and need to have, if we have more than two classes), resulting in many ensemble methods.

Error-Rate of Ensembles

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

$$\hat{h}(x) = \begin{cases} \textit{if } w_1 h_1(x) + \ldots + w_k h_k(x) \geq 0 & : x \mapsto 1 \\ \textit{if } w_1 h_1(x) + \ldots + w_k h_k(x) < 0 & : x \mapsto -1 \end{cases}$$

- The error rate of an ensemble depends on the error rate of the base classifiers (ensemble members) and on how many we combine.
- Assuming $\mathrm{err}(h_1) = \ldots = \mathrm{err}(h_k) = err$, the ensemble error follows a binomial distribution (the ensemble is wrong, if at least half of its members are wrong):

$$\overline{err}\,(k, err) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} err^i (1 - err)^{k-i}$$

(relates to Condorcet's Jury theorem [Marquis de Condorcet, 1785])

Error-Rate of Ensembles

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

Error rate of the ensemble, depending on the number of ensemble members (base classifiers), assuming $err = 0.3$:

$$\overline{err}\,(k, 0.3) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} 0.3^i (1-0.3)^{k-i}$$

Error-Rate of Ensembles

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

$$\overline{err}\,(k, err) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} err^i (1 - err)^{k-i}$$

Independence of Errors

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

$$\overline{err}\,(k, err) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} err^i (1 - err)^{k-i}$$

- Note that we require independence of errors for this formula.
- If the errors are not independent, we cannot expect much improvement from the average.

Conclusions

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

Note that:

We observed two necessary conditions for an improvement of the error rate by combining base classifiers into an ensemble:

1. *All base classifiers are accurate.*
2. *The individual base classifiers are diverse.*

- Accuracy is a mild condition — they need to be at least better than random.
- Diversity: there should be no (strong) correlation between the errors.
- Can we optimize both, diversity and accuracy?

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
  - Aspects of Diversity
  - Bias, Variance, Noise
- Ensemble Methods
- Summary

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
  - Aspects of Diversity
  - Bias, Variance, Noise
- Ensemble Methods
- Summary

Reasons for Diversity

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

Without engineering diversity artificially, there are already inherent reasons for getting diverse classifiers on one and the same classification problem:

- statistical variance
- computational variance
- representation problem
- uncertain/noisy target function

Statistical Variance

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- The hypothesis space is too big to be explored based on the limited amount of training examples.
- Combination of several hypotheses reduces the risk of being very far off.

Computational Variance

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- Some learning algorithms cannot guarantee to find the best hypothesis, as that would be computationally infeasible.
- Instead, they use learning heuristics such that the search could get stuck in local optima.
- Combination of several hypotheses reduces the risk to stick to the wrong (local) optimum.

Representation Problem

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- The hypothesis space does not contain any close approximation to the true target function f.
- The combination of several hypotheses can effectively enlarge the hypothesis space.

Uncertain Target Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- The training examples do not allow to draw unambiguous conclusions on the target function.
  - Noisy training data: there could be contradictory examples.
  - Some class labels might be non-deterministic.
- Combination of several hypotheses reduces the risk, to approximate the wrong target.

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
  - Aspects of Diversity
  - Bias, Variance, Noise
- Ensemble Methods
- Summary

Components of the Error

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

An example from ballistics illustrates different components of the error:

- variance: depends on the force in firing
- noise: extension of the target (it is not a single point)
- bias: depends on the angle α
- $err = bias + variance + noise$

based on a figure by Tan et al. [2006]

Bias, Variance, and Noise in Classification

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

Variance: Depending on variations in training data (e.g., in different folds) or in the choice of parameters or initialization (e.g., different starting weights in a neural net), different hypotheses will be learned.

Noise: The class assignments in the training data are indeterministic or in some other way ambiguous (e.g., there are contradicting training examples).

Bias: A given learning algorithm has certain assumptions on the concept to learn (e.g., on how decision boundaries can be defined).

Note that:

Learning without any assumptions can only result in learning by heart, not in deriving generalizable concepts. "Bias-free learning is futile." [Mitchell, 1997]

Bias vs. Variance

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

Loss Function

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- Given a training set $\mathcal{D} = \{x_i, f(x_i)\}_{i=1}^n$, comprising n points $x_i \in \mathbb{R}^d$ with their corresponding classes $f(x_i)$, a learned hypothesis (classification model) h predicts the class for a given test point x.
- Performance measures (cf. slides 226ff.) focus on minimizing the prediction error in terms of the fraction of misclassified points.
- We can associate costs with making wrong predictions, specifying a *loss function* (with $I : true \mapsto 1, false \mapsto 0$):

  zero-one loss

  $$L(f(x), h(x)) = I(h(x) \neq f(x)) = \begin{cases} 0 & \text{if } h(x) = f(x) \\ 1 & \text{if } h(x) \neq f(x) \end{cases}$$

  squared loss (assuming discrete-valued, not categorical classes) $L(f(x), h(x)) = (f(x) - h(x))^2$

Expected Loss

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

The optimal classifier is the one that minimizes the loss function.

- The true class is not known for a test case x.
- Given posterior class probabilities, we can formalize the expected loss:

$$E_{c \in C}[L(c, h(x))|x] = \sum_{c \in C} L(c, h(x)) \cdot \Pr(c|x)$$

- The goal for the learning model can then be cast as minimization of the expected loss.

Example: Minimizing Expected Zero-One-Loss

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- Using the zero-one loss, minimizing the loss corresponds to minimizing the error rate:

$$\begin{aligned} E_{c\in C}[L(c,h(x))|x] &= E_{c\in C}\left[I(c \neq h(x))|x\right] \\ &= \sum_{c\in C} I(c \neq h(x)) \cdot \Pr(c|x) \\ &= \sum_{c\in C, c\neq h(x)} \Pr(c|x) \\ &= 1 - \Pr(h(x)|x) \end{aligned}$$

- Choose $h(x)$ to maximize the posterior probability, i.e., $h(x) = \arg\max_{c\in C} \Pr(c|x)$ (cf. slide 291).

Bias-Variance Decomposition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

The expected *squared loss* is useful for analyzing the contribution of bias and variance to the error, given some training set $\mathcal{D}$:

$$
\begin{aligned}
&E_C\big[L\big(c, h_{\mathcal{D}}(x)\big) \mid x, \mathcal{D}\big] = E_C\big[\big(c - h_{\mathcal{D}}(x)\big)^2 \mid x, \mathcal{D}\big] \\
&= E_C\big[\big(c \underbrace{- E_C\big[c \mid x\big] + E_C\big[c \mid x\big]}_{\text{add and subtract the same term}} - h_{\mathcal{D}}(x)\big)^2 \mid x, \mathcal{D}\big] \\
&= E_C\Big[\big(c - E_C\big[c \mid x\big]\big)^2 \mid x, \mathcal{D}\Big] + E_C\Big[\big(h_{\mathcal{D}}(x) - E_C\big[c \mid x\big]\big)^2 \mid x, \mathcal{D}\Big] \\
&\quad + E_C\Big[2(c - E_C[c \mid x]) \cdot \big(E_C[c \mid x] - h_{\mathcal{D}}(x)\big) \mid x, \mathcal{D}\Big] \\
&= E_C\Big[\big(c - E_C\big[c \mid x\big]\big)^2 \mid x, \mathcal{D}\Big] + \big(h_{\mathcal{D}}(x) - E_C\big[c \mid x\big]\big)^2 \\
&\quad + 2\big(E_C[c \mid x] - h_{\mathcal{D}}(x)\big) \cdot \underbrace{\big(E_C[c \mid x] - E_C[c \mid x]\big)}_{=0} \\
&= \underbrace{E_C\Big[\big(c - E_C\big[c \mid x\big]\big)^2 \mid x, \mathcal{D}\Big]}_{\mathrm{Var}(C|x)} + \underbrace{\Big(h_{\mathcal{D}}(x) - E_C\big[c \mid x\big]\Big)^2}_{\text{squared error}}
\end{aligned}
$$

Bias-Variance Decomposition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- $\mathrm{Var}(C \mid x) = E_C\left[\left(c - E_C[c \mid x]\right)^2 \mid x, \mathcal{D}\right]$: variance over the classes, given some object x
- $\left(h_{\mathcal{D}}(x) - E_C[c \mid x]\right)^2$: squared error between predicted value $h_{\mathcal{D}}(x)$ and expected value $E_C[c \mid x]$ – depends on the training set.
- Eliminate dependency on training set by averaging over all training sets of the same size.

Bias-Variance Decomposition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

$$
\begin{aligned}
& E_{\mathcal{D}}\Big[\Big(h_{\mathcal{D}}(x) - E_C\big[c \mid x\big]\Big)^2\Big] \\
&= E_{\mathcal{D}}\Big[\Big(h_{\mathcal{D}}(x) \underbrace{-E_{\mathcal{D}}[h_{\mathcal{D}}(x)] + E_{\mathcal{D}}[h_{\mathcal{D}}(x)]}_{\text{add and subtract the same term}} - E_C\big[c \mid x\big]\Big)^2\Big] \\
&= E_{\mathcal{D}}\Big[\big(h_{\mathcal{D}}(x) - E_{\mathcal{D}}[h_{\mathcal{D}}(x)]\big)^2\Big] + E_{\mathcal{D}}\Big[\big(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_C[c \mid x]\big)^2\Big] \\
&\quad + 2\big(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_C[c \mid x]\big) \cdot \underbrace{\big(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_{\mathcal{D}}[h_{\mathcal{D}}(x)]\big)}_{=0} \\
&= \underbrace{E_{\mathcal{D}}\Big[\big(h_{\mathcal{D}}(x) - E_{\mathcal{D}}[h_{\mathcal{D}}(x)]\big)^2\Big]}_{\text{variance}} + \underbrace{E_{\mathcal{D}}\Big[\big(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_C[c \mid x]\big)^2\Big]}_{\text{bias}}
\end{aligned}
$$

Interpretation of the Bias- and Variance-Components of the Classification Error

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

variance: deviation among the learned decision boundaries over different training sets

$$E_{\mathcal{D}}\left[\left(h_{\mathcal{D}}(x) - E_{\mathcal{D}}[h_{\mathcal{D}}(x)]\right)^2\right]$$

bias: systematic deviation of the predicted decision boundary of some classifier from the true decision boundary over different training sets

$$E_{\mathcal{D}}\left[\left(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_{C}[c \mid x]\right)^2\right]$$

Bias, Variance, and Noise — Decomposition

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

Putting together the results from slides 696 and 698, we get the decomposition of the expected squared error into noise, bias, and variance:

$$
\begin{aligned}
& E_{x,\mathcal{D},C}\left[\left(c - h_{\mathcal{D}}(x)\right)^2\right] \\
&= E_{x,\mathcal{D},C}\left[\left(c - E_C[c \mid x]\right)^2 \mid x, \mathcal{D}\right] + E_{x,\mathcal{D}}\left[\left(h_{\mathcal{D}}(x) - E_C[c \mid x]\right)^2\right] \\
&= \underbrace{E_{x,C}\left[\left(c - E_C[c \mid x]\right)^2\right]}_{\text{noise}} \\
&\quad + \underbrace{E_{x,\mathcal{D}}\left[\left(h_{\mathcal{D}}(x) - E_{\mathcal{D}}[h_{\mathcal{D}}(x)]\right)^2\right]}_{\text{average variance}} \\
&\quad + \underbrace{E_{x}\left[\left(E_{\mathcal{D}}[h_{\mathcal{D}}(x)] - E_C[c \mid x]\right)^2\right]}_{\text{average bias}}
\end{aligned}
$$

Bias, Variance, and Noise — Interpretation

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

The expected squared loss over all test points x and training sets $\mathcal{D}$ can therefore be decomposed into three terms:

noise: average variance $\mathrm{Var}(C \mid x)$ over all test points x

- contributes a fixed cost to the loss, independent of the model
- can therefore be ignored when we compare different classifiers

average bias: classifier specific

- indicates whether the model is correct or incorrect
- reflects our assumptions on the domain and adequate decision boundary

average variance: classifier specific

- inconsistency of the classifier decisions
- instability over different training sets

Bias-Complexity Trade-off

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- We could optimize (minimize) the bias (systematic deviation from the optimal decision boundary, i.e., approximation error), if we knew which kind of model fits best to the domain.
- If we don't know this — should we choose some learner with a generally weak bias?
- Learners with a weaker bias are typically producing a more complex model and this way approximate better — on the danger of overfitting.
- Overfitting is a manifestation of a large variance component of the error — the estimation error.
- trade-off between bias and variance:
  - the weaker the bias, the larger the variance
  - the stronger the bias, the smaller the variance

Different Approximation Error of Different Classifiers

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

True decision boundary (dotted) and average (over 100 classifiers trained on 100 variants of the data set) decision boundary (solid) of decision trees (left) and k-nearest neighbor classifiers (right).

based on a figure by Tan et al. [2006]

Different Approximation Error of Decision Trees of Different Depth

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

(a) Decision tree T_1

(b) Decision tree T_2

based on a figure by Tan et al. [2006]

- T_1 and T_2 are trained on the same data.
- T_2 is a pruned version of T_1.
- T_2 has stronger assumptions on separability of classes, i.e., stronger bias.

Bias and Variance of SVM

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

Varying the regularization parameter C for a quadratic kernel (trade-off between slack variables and margin, cf. slide 644):

- small C emphasizes margin (stronger bias)
- large C de-emphasizes margin (weaker bias)

Figure by Zaki and Meira Jr. [2014], cf. their example 22.12.

Ensembles

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Aspects of Diversity
Bias, Variance, Noise
Ensemble Methods
Summary
References

- Any individual classifier would have either a strong bias or a large variance on a non-trivial learning task.
- The combination of classifiers can reduce both, bias and variance:
  - We can combine classifiers with a weak bias, thus a large variance.
  - Averaging reduces the overall variance.

  - We can combine classifiers with strong bias (and thus typically small variance), but choose them in a way to diversify the biases.
  - Averaging reduces the overall bias.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Possibilities to Achieve Diverse Classifiers

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- vary the training set
  - bagging
  - boosting
- manipulate data descriptors
  - use different subspaces/projections
  - use different representations
- manipulate class labels
  - different mappings from polytomous to dichotomous problems

} meta methods

- manipulate learning algorithm
  - use elements of randomness
  - use different start configurations for local optimizers

} specialized methods

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Varying the Training Set

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- based on the instability of learning algorithms:
  - An algorithm is the more stable, the less classifiers (hypotheses) differ that have been learned on varied training data for the same classification problem.
  - Instability is based on the variance of the learned decision boundary: high variance makes the learner susceptible to overfitting.
  - For an unstable learning algorithm, small changes of the training set can induce large changes in the model.
- To build ensembles based on varying the training set, instable learners are beneficial, e.g.:
  - decision trees
  - neuronal nets
  - rule learners (not covered in the lecture)

Bagging

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- Bagging is an acronym for **B**ootstrap **Agg**regat**ing**.
- Idea: get diverse training sets by repeated bootstrapping (cf. slide 223).

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Bagging trains a classifier on each bootstrap sample and aggregates the models.
- With unstable learners, sufficiently diverse hypotheses will be learned.
- New data objects are classified by voting over all learned hypotheses.

Boosting

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- While the bootstrap is sampled uniformly, boosting assigns a weight to each data object.
- The weights are adjusted (increased) for difficult objects (where previous hypotheses made errors).
- The weights change the probability of drawing the object in the next round of sampling.
- As a result, difficult objects will show up more frequently in the next round and thus get implicitely a higher weight for training the classifier.

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

Recommended Reading:

For an extensive discussion of boosting, see the book by Schapire and Freund [2012].

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Varying the Data Descriptors

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- "feature bagging":
  - sample attributes
  - learn in the subspace
  - repeat several times and combine the models/predictions
- instead of sampling individual attributes, use different feature combinations/projections (e.g., random projections [Johnson and Lindenstrauss, 1984, Achlioptas, 2001])
- use different feature spaces [Kittler et al., 1998]

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Manipulating Class Labels

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- Some classification algorithms can only solve dichotomous classification problems (e.g., SVMs).
- Complex problems with more than 2 classes can be tackled by learning several classifiers on subproblems, reduced to two classes.
- Most prominent methods:
  - one-versus-rest
  - all pairs
  - Error correcting output codes (ECOC)

One-Versus-Rest

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- a.k.a. one-versus-all, one-versus-others, one-per-class
- For n classes, we train n classifiers, each separating one of the classes, in turn, from all the others.

Figures by Fürnkranz [2002].

All Pairs

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- a.k.a. all-versus-all, one-versus-one, round robin, pairwise
- For each pair of n classes, we train a classifier for separating just these classes from each other.

Figures by Fürnkranz [2002].

Combination of Multiple Two-Class Classifiers

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- collect all votes
- directed acyclic graph [Platt et al., 1999]: sequential votes, follow only up on the winners

Combination of Multiple Two-Class Classifiers

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- nested dichotomies [Frank and Kramer, 2004]

- incorporate domain knowledge (hierarchies of classes): hierarchically nested dichotomies [Zimek et al., 2010]

Error Correcting Output Codes (ECOC)

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

ECOC defines both: getting diverse classifiers, and combining their votes.

Diversity

- Set C of classes is randomly split k times into two subsets A and B.
- Examples $\in A$ get assigned the new label -1, the other classes ($\in B$) get label 1.
- Train k classifiers on the resulting k two-class problems.

Combination

- If in one of the problems a classifier votes for A, all classes $\in C$ that belong to A in this iteration get a vote.
- The class $\in C$ receiving most votes is the decision of the ensemble.

ECOC — Example

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- Let $C = \{c_1, c_2, c_3, c_4\}$, choose $k = 7$ (i.e., we have a 7-bit encoding):

class	code						
c_1	1	1	1	1	1	1	1
c_2	0	0	0	0	1	1	1
c_3	0	0	1	1	0	0	1
c_4	0	1	0	1	0	1	0

- For each bit of the code, we train a classifier.
- We get seven decisions, e.g., $(0, 1, 1, 1, 1, 1, 1)$ — what is the ensemble's decision?

ECOC — Motivation

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- The name "error correcting" relates to the idea that we introduce a redundancy for the decisions.
- The codes can be chosen randomly.
- For a good diversity, the codes should separate well:

  row separation: each pair of codes should have a large Hamming-distance.

  column separation: the k binary classifiers should be rather uncorrelated.
- What is the Hamming distance of the vote $(0, 1, 1, 1, 1, 1, 1)$ in our example?

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Use vs. Design of Random Elements

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

- Some learning algorithms include random elements.
- An example are random starting points for local optimizers (e.g., the weights used in a neural net).
- Other learning algorithms can be changed to incorporate random elements.
- The greedy optimization used in decision trees is an obvious choice, leading to random forests.

Random Forests

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

"Random Forests" is a speaking name for an ensemble consisting of trees — decision trees that involve some random component:

- train each decision tree on an independent randomly selected subset of the features (Forest-RI: Random Input);
- generate at each node a set of random linear combinations of a subset of the features, select the best of them for the split (Forest-RC: Random Combination);
- select at each node randomly one of the n best splits;

  or

- some combination of the three approaches.

Outline

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
  - Varying the Training Set
  - Varying Data Descriptors
  - Manipulating Class Labels
  - Manipulating the Learning Algorithm
  - Domain-Specific Ensembles: Learn to Combine
- Summary

Methods for Subcellular Localization Prediction

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Example: Decision Tree for Combination

DM566/DM868/DM870
Arthur Zimek
Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Varying the Training Set
Varying Data Descriptors
Manipulating Class Labels
Manipulating the Learning Algorithm
Domain-Specific Ensembles: Learn to Combine
Summary
References

Methods for prediction of subcellular localization of proteins are divers, based on different principles ($\Longrightarrow$ different bias).

- Combine them by voting.
- Specialized combination: train a decision tree to predict when to trust which predictor most [Aßfalg et al., 2009].

```
cello = 1:  1
cello = 2
| ploc = 1:  1
| ploc = 2
| | ipsort = n:  2
| | ipsort = s
| | | wolf = 1:  3
| | | wolf = 2:  2
⋮
```

```
⋮
| | multi = 11:  1
| ploc = 8:  2
cello = 3:  3
cello = 4:  1
cello = 6
| wolf = 1
| | predotar = n:  6
⋮
```

Outline

Introduction
Frequent Pattern Mining
Feature Spaces
Clustering – Basics and k-means
Classification – Basics and a Basic Classifier
Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Distributions and Learning with Distributions
Entropy, Purity, and Separation: Linear vs. Non-Linear Separation
Ensemble Learning
- Motivation for Ensemble Learning
- Diversity
- Ensemble Methods
- Summary

Summary

DM566/DM868/DM870

Arthur Zimek

Introduction
Freq. Pattern Mining
Feature Spaces
Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with Distributions
Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning
Motivation for Ensemble Learning
Diversity
Ensemble Methods
Summary
References

You learned in this section:

- *accuracy and diversity of ensemble members and their relation to theoretical expectations for ensemble quality*
- *sources of diversity*
- *bias-variance-decomposition*
- *bias-complexity trade-off*
- *relation of bias and variance to overfitting and instability*
- *methods to achieve diverse ensemble members*
- *ensemble combination methods*

References I

D. Achlioptas. Database-friendly random projections. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA*, 2001.

C. C. Aggarwal and J. Han, editors. *Frequent Pattern Mining*. Springer, 2014.

C. C. Aggarwal and C. K. Reddy, editors. *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Helsinki, Finland*, pages 15–26, 2002. doi: 10.1007/3-540-45681-3_2.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Philadelphia, PA*, pages 49–60, 1999. doi: 10.1145/304182.304187.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, LA*, pages 1027–1035, 2007. doi: 10.1145/1283383.1283494.

References II

J. Aßfalg, J. Gong, H.-P. Kriegel, A. Pryakhin, T. Wei, and A. Zimek. Supervised ensembles of prediction methods for subcellular localization. *Journal of Bioinformatics and Computational Biology*, 7(2):269–285, 2009. doi: 10.1142/S0219720009004072.

V. Barnett. The study of outliers: Purpose and model. *Applied Statistics*, 27(3): 242–250, 1978.

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek. Quality of similarity rankings in time series. In *Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD), Minneapolis, MN*, pages 422–440, 2011. doi: 10.1007/978-3-642-22922-0_25.

F. Birzele and S. Kramer. A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, 22(21):2628–2634, 2006. doi: 10.1093/bioinformatics/btl453.

M. M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 93–104, 2000. doi: 10.1145/342009.335388.

References III

R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Gold Coast, Australia*, pages 160–172. 2013a. doi: 10.1007/978-3-642-37456-2_14.

R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery*, 27(3):344–371, 2013b. doi: 10.1007/s10618-013-0311-4.

R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):5:1–51, 2015. doi: 10.1145/2733381.

G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30:891–927, 2016. doi: 10.1007/s10618-015-0444-8.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

D. Defays. An efficient algorithm for the complete link cluster method. *The Computer Journal*, 20(4):364–366, 1977. doi: 10.1093/comjnl/20.4.364.

References IV

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–31, 1977.

J. L. Devore and K. N. Berk. *Modern Mathematical Statistics with Applications*. Springer, 2nd edition, 2012.

M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 3rd edition, 2009. ISBN 9783662443415.

T. G. Dietterich. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems (MCS), Cagliari, Italy*, pages 1–15, 2000. doi: 10.1007/3-540-45014-9_1.

T. G. Dietterich. Ensemble learning. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 405–408. MIT Press, 2nd edition, 2003.

P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the 13th International Conference on Machine Learning (ICML), Bari, Italy*, pages 105–112, 1996.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley&Sons, 2nd edition, 2001.

References V

C. Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th International Conference on Machine Learning (ICML), Washington, DC*, pages 147–153, 2003.

L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM), San Francisco, CA*, 2003. doi: 10.1137/1.9781611972733.5.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pages 226–231, 1996.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pages 82–88, 1996.

U. M. Fayyad, C. Reina, and P. S. Bradley. Initialization of iterative refinement clustering algorithms. In *Proceedings of the 4th ACM International Conference on Knowledge Discovery and Data Mining (KDD), New York City, NY*, pages 194–198, 1998.

References VI

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

E. Frank and S. Kramer. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning (ICML), Banff, AB, Canada*, 2004. doi: 10.1145/1015330.1015363. URL `http://doi.acm.org/10.1145/1015330.1015363`.

J. Fürnkranz. Round robin classification. 2:721–747, 2002.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2):1–12, 2000. doi: 10.1145/342009.335372.

J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

J. A. Hartigan. *Clustering Algorithms*. John Wiley&Sons, New York, London, Sydney, Toronto, 1975.

References VII

V. Hautamäki, I. Kärkkäinen, and P. Fränti. Outlier detection using k-nearest neighbor graph. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Cambridge, England, UK*, pages 430–433, 2004. doi: 10.1109/ICPR.2004.1334558.

D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany*, pages 482–500, 2010. doi: 10.1007/978-3-642-13818-8_34.

Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Tucson, AZ*, 1997.

R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, C-22(11): 1025–1034, 1973.

W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore*, pages 577–593, 2006. doi: 10.1007/11731139_68.

References VIII

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

E. M. Knorr and R. T. Ng. A unified notion of outliers: Properties and computation. In *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, CA*, pages 219–222, 1997. doi: 10.1145/782010.782021.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, pages 1649–1652, 2009. doi: 10.1145/1645953.1646195.

H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3): 231–240, 2011. doi: 10.1002/widm.30.

References IX

H.-P. Kriegel, E. Schubert, and A. Zimek. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems (KAIS)*, 52(2):341–378, 2017. doi: 10.1007/s10115-016-1004-2.

L. J. Latecki, A. Lazarevic, and D. Pokrajac. Outlier detection with kernel density functions. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), Leipzig, Germany*, pages 61–75, 2007. doi: 10.1007/978-3-540-73499-4_6.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982. doi: 10.1109/TIT.1982.1056489.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics*, volume 1, pages 281–297, 1967.

H. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander. On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management (SSDBM), San Diego, CA*, pages 7:1–12, 2015. doi: 10.1145/2791347.2791352.

M. J. A. N. C. Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, Paris, 1785.

References X

L. McInnes and J. Healy. Accelerated hierarchical density based clustering. In *ICDM Workshops*, pages 33–42, 2017.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

M. Mitzenmacher and E. Upfal. *Probability and Computing. Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.

S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In *Proceedings of the 19th International Conference on Data Engineering (ICDE), Bangalore, India*, pages 315–326, 2003. doi: 10.1109/ICDE.2003.1260802.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), Denver, CO*, pages 547–553, 1999.

References XI

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi: 10.1023/A:1022643204877.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 2014. doi: 10.1109/TKDE.2014.2365790.

S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 427–438, 2000. doi: 10.1145/342009.335437.

D. M. Raup. Geometric analysis of shell coiling: General problems. *Journal of Paleontology*, 40(5):1178–1190, 1966.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

P. J. Rousseeuw and M. Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79, 2011. doi: 10.1002/widm.2.

J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998. doi: 10.1023/A:1009745219419.

References XII

R. E. Schapire and Y. Freund. *Boosting. Foundations and Algorithms*. MIT Press, Cambridge, MA, 2012.

E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA*, pages 1047–1058, 2012. doi: 10.1137/1.9781611972825.90.

E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014a. doi: 10.1007/s10618-012-0300-z.

E. Schubert, A. Zimek, and H.-P. Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, pages 542–550, 2014b. doi: 10.1137/1.9781611973440.63.

B. Schölkopf and A. J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning. From Theory to Algorithms*. Cambridge University Press, 2014.

References XIII

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973. doi: 10.1093/comjnl/16.1.30.

R. Silberzahn and E. L. Uhlmann. Many hands make tight work. *Nature*, 526: 189–191, 2015. doi: 10.1038/526189a.

B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1986.

R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Montreal, QC, Canada*, pages 1–12, 1996. doi: 10.1145/233269.233311.

N. N. Taleb. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House, 2001.

N. N. Taleb. *The Black Swan. The Impact of the Highly Improbable*. Random House, 2007.

P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

References XIV

J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan*, pages 535–548, 2002. doi: 10.1007/3-540-47887-6_53.

G. Valentini and F. Masulli. Ensembles of learning machines. In *Proceedings of the 13th Italian Workshop on Neural Nets, Vietri, Italy*, pages 3–22, 2002. doi: 10.1007/3-540-45808-5_1.

L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235, 2010. doi: 10.1002/sam.10080.

J. Vreeken and N. Tatti. Interesting patterns. In C. C. Aggarwal and J. Han, editors, *Frequent Pattern Mining*, chapter 5, pages 105–134. Springer, 2014. doi: 10.1007/978-3-319-07821-2_5.

I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 3rd edition, 2011.

M. Zaki and W. Meira Jr. *Data Mining and Analysis. Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

References XV

K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, pages 813–822, 2009. doi: 10.1007/978-3-642-01307-2_84.

A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1280, 2018. doi: 10.1002/widm.1280.

A. Zimek, F. Buchwald, E. Frank, and S. Kramer. A study of hierarchical and flat classification of proteins. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):563–571, 2010. doi: 10.1109/TCBB.2008.104.

A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012. doi: 10.1002/sam.11161.

A. Zimek, I. Assent, and J. Vreeken. Frequent pattern mining algorithms for data clustering. In C. C. Aggarwal and J. Han, editors, *Frequent Pattern Mining*, chapter 16, pages 403–423. Springer, 2014. doi: 10.1007/978-3-319-07821-2_16.